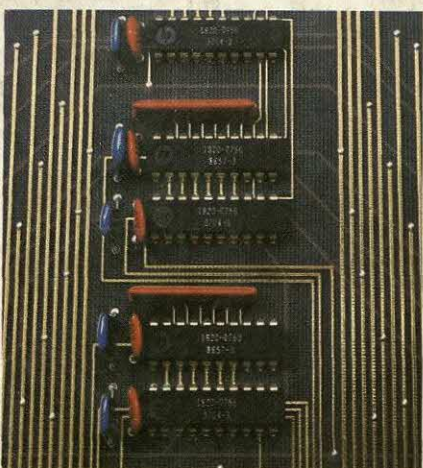
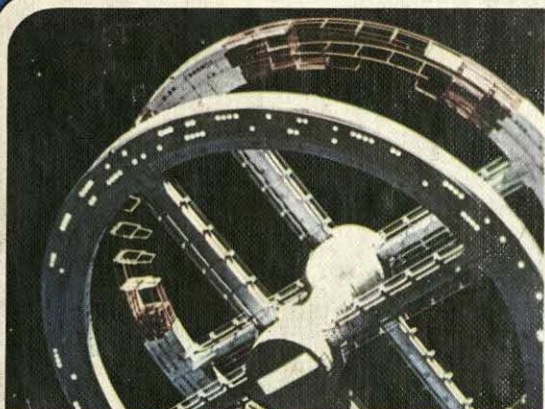


1

THE NEW BOOK OF POPULAR SCIENCE



THE NEW BOOK OF POPULAR SCIENCE

EDITORIAL

Editorial Director: Benjamin S. Chao

Executive Editor: Lynn Grace Miller

Editor: David B. Lachner, Art Director: Walter R. M. ...

Editor

Robert S. Anderson, George T. ...

Harold Kinde, John ...

Steven ...

Stanley ...

88-81-28-120
SIEH

THE NEW BOOK OF POPULAR SCIENCE

VOLUME 1

Astronomy & Space Science
Computers & Mathematics



P.O.E.R.T., West Bengal.
Date 26-12-88
Acc. No. 4313

COPYRIGHT © 1984 BY



Copyright © 1982, 1981, 1980, 1979, 1978 by GROLIER INCORPORATED

Copyright © Philippines 1979 by GROLIER INTERNATIONAL, INC.

Copyright © Republic of China 1978 by GROLIER INTERNATIONAL, INC.

Library of Congress Cataloging in Publication Data

Main entry under title:
The New book of popular science.

Includes bibliographies and index.

Contents: v. 1. Astronomy & space science.

Computers & mathematics.

1. Science—Popular works—Collected works.
2. Technology—Popular works—Collected works.
3. Natural history—Popular works—Collected works.

Q162.N437 1984 500 83-26595

ISBN 0-7172-1213-0 (set)

No part of THE NEW BOOK OF POPULAR SCIENCE may be reproduced without special permission in writing from the publishers

PRINTED
IN
U.S.A.

THE NEW BOOK OF POPULAR SCIENCE

EDITORIAL

Editorial Director Bernard S. Cayne

Executive Editor Lynn Giroux Blum

Managing Editor Doris E. Lechner **Art Director** Walter A. Schwarz

Editors

Robert S. Anderson Barbara Tchabovsky

Herbert Kondo John Tedford

Steven Moll Jenny Tesar

Stanley Schindler Linda Triegel

Chief Indexer Jill Schuler **Indexer** Alexis Kasden

Production Editor Nicole A. Vanasse **Staff Assistant** Jennifer Drake

Proofreader Stephen Romanoff

Editorial Financial Manager S. Jean Gianazza **Librarian** Charles Chang

Picture Research

Director, Central Picture Services John Schultz

Manager, Picture Library Jane H. Carruth **Head, Photo Research** Ann Eriksen

Photo Researchers A.E. Raymond Elissa G. Spiro

Assistant Photo Librarian Mickey Austin

MANUFACTURING

Director of Manufacturing Harriet Ripinsky

Senior Production Manager Joseph J. Corlett

Production Manager Alan Phelps

Production Assistant Marilyn Smith

CONTRIBUTORS

ELISABETH ACHELIS, *Former President, World Calendar Association, Inc.*

THE CALENDAR

JOHN G. ALBRIGHT, Ph.D., *Former Chairman, Department of Physics, University of Rhode Island*

HEAT TRANSMISSION

FIKRI ALICAN, M.D., *Assistant Professor of Surgery, The University of Mississippi Medical Center, Jackson, Mississippi*

coauthor, ORGAN TRANSPLANTS

LAWRENCE K. ALTMAN, M.D., *Medical Reporter, New York Times*

INFLUENZA

STANLEY W. ANGRIST, Ph.D., *Department of Mechanical Engineering, Carnegie-Mellon University*

ENTROPY

EDWARD V. APPLETON, D.Sc., *Nobel Prize winner in Physics (1947)*

RADIO ASTRONOMY

ELIAS M. AWAD, M.B.A., M.A., *Associate Professor, Graduate School of Business, DePaul University*

DATA PROCESSING

GERALD AXELROD, *Future School*

PERSONAL COMPUTERS

S. HOWARD BARTLEY, Ph.D., *Professor of Psychology, Michigan State University*

THE EAR AND HEARING

THE EYE AND VISION

J. FREMONT BATEMAN, M.D., *Former Superintendent, Columbus State Hospital, Columbus, Ohio*

THE NERVOUS SYSTEM

W. W. BAUER, M.D., *Former consultant in health education; former Director, Department of Health Education, American Medical Association*

DEFENSES OF THE BODY

J. KELLY BEATTY, *Staff Member, Sky and Telescope Corporation*

COMMUNICATIONS SATELLITES

R. F. BECKWITH, *Advertising Manager, Recordak Corporation*

MICROFILM

HANS J. BEHM, *Science writer*

THE EXPLORATION OF THE MOON

THEODORE D. BENJAMIN, A.M., *Head of Science Department, DeWitt Clinton High School, New York City*

ENGINES

LYMAN BENSON, *Professor Emeritus of Botany, Pomona College; author, Cacti of the United States and Canada*

CACTUS PLANTS

M. H. BERRY, M.S., *Chairman, Division of Science and Mathematics, and Professor of Botany, West Liberty State College*

ALGAE

CHRISTOPHER J. BISE, Ph.D., *Assistant Professor of Mining Engineering, College of Earth & Mineral Sciences, Pennsylvania State University*

COAL

LOUIS FAUGERES BISHOP, M.D., *Cardiologist; visiting physician, Bellevue Hospital, New York City*

BLOOD

NICHOLAS T. BOBROVNIKOV, Ph.D., *Professor Emeritus, Astronomy, Ohio State University; former Director, Perkins Observatory*

COMETS

BART J. BOK, Ph.D., *Chairman, Astronomy Department, University of Arizona*

TIDES

YVONNE BONNAFOUS, *Science writer*

THE PRIMITIVE CHORDATES

FRANKLYN M. BRANLEY, Ph.D., *Astronomer, American Museum of Natural History—Hayden Planetarium*

THE EARTH

FRANK A. BROWN, Jr., Ph.D., *Morrison Professor of Biology, Northwestern University*

BIOLOGICAL RHYTHMS AND CLOCKS

KEITH E. BULLEN, *Science writer*

THE INTERIOR OF THE EARTH

ALAN C. BURTON, Ph.D., *Professor of Biophysics, The University of Western Ontario Faculty of Medicine*

EXERCISE AND REST

JOHN H. CALLENDER, B.A., B.Arch., *Professor of Architecture, Pratt Institute*

BUILDING TECHNOLOGY

H. A. CATES, M.B., *Former Professor of Anatomy, University of Toronto*

THE BONES OF THE BODY

MORRIS CHAFETZ, M.D., *Principal Research Scientist, The Johns Hopkins University; former Director, National Institute of Alcohol Abuse and Alcoholism*

ALCOHOLISM

MARK CHARTRAND III, Ph.D., *Director, Hayden Planetarium*

PLANETARIUMS

JANE CHESNUTT, *Editor, Environmental Information Center*

WILDERNESS

CLYDE M. CHRISTENSEN, Ph.D., *Professor of Plant Pathology, University of Minnesota*

LEAVES

ROOTS AND STEMS

GEORGE L. CLARK, Ph.D., *Research Professor of Chemistry, University of Illinois*

MICROSCOPES

DANIEL M. COHEN, Ph.D., *Laboratory Director, Ichthyological Laboratory, Bureau of Commercial Fisheries, U.S. Fish and Wildlife Service*

FISHES

JAMES S. COLES, Ph.D., *President, Bowdoin College*

SOLUTIONS

IAN McT. COWAN, Ph.D., *Dean, Faculty of Graduate Studies; Prof. of Zoology, University of British Columbia*

MAMMALS

JOSEPH G. COWLEY, *Research Institute of America*

GAME THEORY

WILLIAM J. CROMIE, *Executive Director, Council for the Advancement of Science Writing*

ALTERNATE ENERGY SOURCES

F. JOE CROSSWHITE, Ph.D., *Associate Professor of Mathematics Education, Ohio State University*

PROBABILITY

STATISTICS

G. EDWARD DAMON, *Former Consumer Writer, Bureau of Foods, U.S. Food and Drug Administration*

A PRIMER OF CARBOHYDRATES, FATS,

AND MINERALS

A PRIMER OF PROTEINS AND FIBER

GÉRARD De VAUCOULEURS, Ph.D., *Professor, Department of Astronomy, University of Texas*

THE GALAXIES

THEODOSIUS DOBZHANSKY, D.Sc., *Former Professor, Rockefeller University*

HUMAN HEREDITY

ROY DUBISCH, Ph.D., *Professor, Department of Mathematics, University of Washington*

SET THEORY

MAC V. EDDS, Jr., Ph.D., *Chairman, Division of Medical Science, Brown University*

THE SKIN

STRUCTURE OF THE BODY

HOWARD F. FEHR, Ph.D., *Professor of Mathematics, Head of the Department of the Teaching of Mathematics, Teachers College, Columbia University*

ANALYTIC GEOMETRY

ARITHMETIC

others

GERALD FEINBERG, Ph.D., Professor, Department of Physics, Columbia University

ELEMENTARY PARTICLES

IRWIN K. FEINSTEIN, Ph.D., Instructor, Mathematics, University of Illinois at Chicago Circle

THE BINARY NUMERAL SYSTEM

E. P. FELCH, M.E., Director, Missile and Magnetic Apparatus Laboratory, Bell Telephone Laboratories

GUIDANCE IN SPACE

GEORGE B. FIELD, Ph.D., Director, Center for Astrophysics, Harvard University and the Smithsonian Institution

THE ORIGIN OF THE UNIVERSE

F. L. FITZPATRICK, Ph.D., Former Chairman, Department of Natural Sciences, Teachers College, Columbia University

FLATWORMS AND ROUNDWORMS

MOLLUSKS

others

JOHN A. FLEMING, B.S., D.Sc., Former Director of the Department of Terrestrial Magnetism, Carnegie Institution, Washington, D.C.

MAGNETISM

WILLIAM S. FOSTER, B.S. in Ch.E., Editor, The American City

THE WATER SUPPLY

EUGENE E. FOWLER, M.S., Director, Division of Isotopes Development, U.S. Atomic Energy Commission

RADIOISOTOPES

PAUL J. FOX, Ph.D., Lamont-Doherty Geological Observatory, Columbia University

DEEP-SEA EXPLORATION

LAURENCE W. FREDRICK III, Ph.D., Professor of Astronomy and Director, Leander McCormick Observatory, University of Virginia

TELESCOPES

SMITH FREEMAN, M.D., Ph.D., Chairman, Department of Biochemistry, Northwestern University

THE DIGESTIVE SYSTEM

M. A. FREIBERG, Ph.D., Argentina Museum of Natural History

ANIMAL BEHAVIOR

HARRY J. FULLER, Ph.D., Former Professor of Botany, University of Illinois

BACTERIA

SEED AND FRUIT DISPERSAL

others

ELDON J. GARDNER, Science writer

CHARLES DARWIN

A. B. GARRETT, Ph.D., Vice President for Research, Ohio State University

COLLOIDS

ELECTROCHEMISTRY

WATER

CHARLES N. GAYLORD, M.S.E., Professor and Chairman, Department of Civil Engineering, and Assistant Dean, School of Engineering, University of Virginia

MODERN ENGINEERING

CHALMERS L. GEMMILL, M.D., Chairman, Department of Pharmacology, University of Virginia School of Medicine

THE CIRCULATORY SYSTEM

WILLIS J. GERTSCH, Ph.D., Curator of the Department of Entomology, American Museum of Natural History

SPIDERS AND THEIR KIN

BENTLEY GLASS, Ph.D., Vice President for Academic Affairs and Distinguished Professor of Biological Sciences, State University of New York at Stony Brook

GROWTH, DEVELOPMENT, AND DECLINE

PHILLIP GOLDSTEIN, M.S., Chairman, Department of Biology, Abraham Lincoln High School, Brooklyn, New York
HOW TO DO AN EXPERIMENT
HOW TO USE A MICROSCOPE

RICHARD P. GOLDTHWAIT, Ph.D., Chairman, Department of Geology, Ohio State University

GLACIERS

MARVIN R. GORE, Dean, School of Business Administration, Mount San Antonio College

MICROPROCESSORS

DONALD C. GREGG, Ph.D., Pomeroy Professor of Chemistry, University of Vermont

ORGANIC CHEMISTRY

JAMES J. HAGGERTY, Science writer

JET PROPULSION

KATHERINE HARAMUNDANIS, B.A., Supervisor, Star Catalog Project, Smithsonian Astrophysical Observatory

THE MOON

FORREST E. HARDING, D.B.A., Associate Professor of Marketing, California State University

URBAN MASS TRANSPORTATION

JAMES D. HARDY, M.D., Professor and Chairman, Department of Surgery, The University of Mississippi Medical Center, and Surgeon-in-Chief to the University Hospital, Jackson, Mississippi
coauthor, ORGAN TRANSPLANTS

F. R. HARNDEN, Jr., Ph.D., Astrophysicist, Harvard-Smithsonian Center for Astrophysics

X-RAY ASTRONOMY

KENNETH HARWOOD, Ph.D., Department of Telecommunications, University of Southern California

TELEVISION

E. NEWTON HARVEY, Ph.D., Former Professor Emeritus of Biology, Princeton University

BIOLUMINESCENCE

JAMES HASSETT, Ph.D., Department of Psychology, Boston University

PSYCHOSOMATIC MEDICINE

JOEL W. HEDGPETH, Ph.D., Professor of Zoology and Director of Pacific Marine Station, College of the Pacific

THE SEA

LOUIS M. HEIL, Ph.D., Director, Office of Testing and Research, Brooklyn College

GRAVITATION

ANTONY HEWISH, Ph.D., F.R.S., Nobel Prize winner in Physics (1974); Professor of Radio Astronomy, University of Cambridge

PULSARS

NORMAN E. A. HINDS, Ph.D., Former Professor of Geology, University of California

HOW RUNNING WATER CHANGES THE LAND

THOMAS E. HITCHINGS, B.A., Natural Science Editor, Book Development Group, John Wiley & Sons, Inc.

TAPE RECORDERS, CASSETTES, AND CARTRIDGES

LILLIAN HODDESON, Ph.D., Department of Physics, Rutgers University

QUANTUM THEORY

GEORGE HOEFER, B.S. in E.E., former Associate Editor, Electronic Equipment Engineering magazine

RADIO

ROBERT W. HOWE, Ed.D., Assistant Professor of Science Education, Ohio State University

BIRD OBSERVATION

JOHN B. IRWIN, Ph.D., Associate Professor, Department of Earth Sciences, Newark State College

OBSERVATORIES

ALEXANDER JOSEPH, D.Ed., Professor, Department of Physics, Bronx Community College, New York

SIMPLE MACHINES

CORLISS G. KARASOV, Science writer

WASTE DISPOSAL

MORLEY KARE, Ph.D., Professor of Physiology, University of Pennsylvania

BIONICS

JOHN KENTON, Science writer

NUCLEAR ENERGY

HAROLD P. KNAUSS, Ph.D., Former Head, Department of Physics, University of Connecticut

MUSICAL SOUNDS
SOUND

C. ARTHUR KNIGHT, Ph.D., Professor, Molecular Biology; Research Biochemist, University of California, Berkeley

VIRUSES

SERGE A. KORFF, Ph.D., Professor of Physics, New York University

THE THEORY OF RELATIVITY

STEPHEN N. KREITZMAN, Ph.D., Assistant Professor, Dental Research, Emory University

BIOCHEMISTRY
DNA AND RNA

LEONARD C. LABOWITZ, Ph.D., Physical chemist and science writer

SUPERCONDUCTIVITY

MORT LA BREQUE, Editor, The Sciences

BLACK HOLES

DONALD A. LAIRD, Ph.D., Author; former Director, Colgate University Psychological Laboratory

PSYCHOLOGY
SLEEP

E. V. LAITONE, Ph.D., Professor of Aeronautical Sciences, University of California College of Engineering

SUPERSONIC FLIGHT

LYNN LAMOREAUX, Ph.D., MRC Radiobiology Unit, Great Britain

EMBRYOLOGY
GENETIC ENGINEERING

T. K. LANDAUER, Ph.D., Bell Telephone Laboratories

MEMORY

ROBIN LANIER, Free-lance writer

PHONOGRAPH RECORDS AND RECORDING
VIDEOCASSETTES AND VIDEODISCS

HENRY LANSFORD, M.A., National Center for Atmospheric Research

ENVIRONMENTAL POLLUTION
WEATHER

PETER A. LARKIN, D. Phil., Institute of Animal Research Ecology, University of British Columbia

CONSERVATION

THOMAS GORDON LAWRENCE, A.M., Former Chairman, Biology Department, Erasmus Hall High School, Brooklyn, New York

MOSES AND FERNS
TREES

BENEDICT A. LEERBURGER, Science writer

OFFICE AUTOMATION

AARON A. LERNER, M.D., School of Medicine, Yale University

ALBERT EINSTEIN
ISAAC NEWTON

WILLY LEY, Ph.D., Former rocket expert and lecturer

MANNED SPACE FLIGHTS
others

WILLARD F. LIBBY, Ph.D., Nobel Prize winner in Chemistry (1960); Professor Emeritus, Department of Chemistry, UCLA

RADIOCARBON DATING

W. THOMAS LIPPINCOTT, Ph.D., Professor and Academic Vice Chairman, Department of Chemistry, Ohio State University

CHROMATOGRAPHY

M. STANLEY LIVINGSTON, Ph.D., Former Professor of Physics, Massachusetts Institute of Technology; former Director, Cambridge Electron Accelerator, and Associate Director, National Accelerator Laboratory

ATOM SMASHERS

ARMIN K. LOBECK, Ph.D., Former Professor of Geology, Columbia University

CAVES

BARBARA LOBRON, Free-lance writer, photography; former Managing Editor, Camera 35

PHOTOGRAPHY

MARIAN LOCKWOOD, Science writer

THE CONSTELLATIONS
THE NIGHT SKY

SUZANNE LOEBL, Science Editor, The Arthritis Foundation

ARTHRITIS

DONALD B. LOURIA, M.D., Professor and Chairman, Department of Medicine and Community Health, New Jersey Medical School

DRUG ABUSE

PAUL D. LOWMAN, Jr., Ph.D., Aerospace Technologist, Goddard Space Flight Center (NASA)

GEOLOGY OF THE MOON

CHARLES C. MACKLIN, M.D., Former Professor of Histology and Embryology, University of Western Ontario

HUMAN EMBRYOLOGY

HOWARD O. McMAHON, Ph.D., President, Arthur D. Little, Inc.

CRYOGENICS

OSCAR E. MEINZER, Ph.D., Former Chief, Division of Ground Water, United States Geological Survey

WATER IN THE GROUND

HELEN HYNSON MERRICK, Former Senior Editor, Grollier Incorporated

X RAYS

LON W. MORREY, D.D.S., Editor Emeritus, Journal of the American Dental Association

TEETH

WALTER MUNK, Ph.D., Director, La Jolla Unit, and Associate Director, Institute of Geophysics, UCLA

WAVES

CHARLES MERRICK NEVIN, Science writer

CLIMATES OF THE PAST

IAN NISBET, Environmental consultant

ACID RAIN

LINDSAY S. OLIVE, Ph.D., Professor of Botany, Columbia University

FUNGI

JAMES A. OLIVER, Ph.D., Director, New York Zoological Society

LIZARDS

DAMON R. OLSZOWY, M.S., Ph.D., Biology Department, State University of New York at Farmingdale

CLONING

FREDERICK I. ORDWAY III, Professor of Science and Technology, University of Alabama

SPACE STATIONS

IRVING H. PAGE, M.A., Research Associate, U.S. Naval Research Laboratory

RADAR

JOHN C. PALLISTER, Research Associate, Department of Entomology, American Museum of Natural History

ANIMAL MIGRATION

BEEETLES
BUTTERFLIES AND MOTHS
INSECTS

- CECILIA PAYNE-GAPOSCHKIN, D.Sc.**, *Smithsonian Astrophysical Observatory, Harvard University*
INTERSTELLAR SPACE
THE MOON
THE STARS
- RICHARD F. POST, Ph.D.**, *Director, Magnetic Mirror Program, Lawrence Radiation Laboratory, University of California*
PLASMAS
- FREDERICK H. POUGH, Ph.D.**, *Consulting mineralogist; former Curator of Minerals and Gems, American Museum of Natural History*
COLLECTING ROCKS AND MINERALS
MINERALS
- TERENCE T. QUIRKE, Ph.D.**, *Former Professor of Geology, University of Illinois*
EARTH'S CRUST
- HOWARD G. RAPAPORT, M.D.**, *Associate Professor of Clinical Pediatrics (Allergy), Albert Einstein College of Medicine, Yeshiva University*
ALLERGIES
- ERIC RODGERS, Ph.D.**, *Dean of the Graduate School, University of Alabama*
PROPERTIES OF MATTER
- EDWARD ROSEN, Ph.D.**, *City University of New York*
COPERNICUS
- ALBERT J. RUF, Science writer
HEAT**
- BETTE RUNCK, Science writer
BIOFEEDBACK**
- J. H. RUSH, Ph.D.**, *Physicist, National Center for Atmospheric Research, Boulder, Colorado*
COLOR
OPTICS
- FRANCIS J. RYAN, Ph.D.**, *Former Professor of Biology, Columbia University*
coauthor, SPONGES AND COELENTERATES
STARFISH AND OTHER ECHINODERMS
- JOHN D. RYDER, Ph.D.**, *Dean, College of Engineering, Michigan State University*
AN INTRODUCTION TO ELECTRONICS
- BETH SCHULTZ, Ed.D.**, *Department of Biology, Western Michigan University*
ECOLOGY
PHOTOSYNTHESIS
- GARY E. SCHWARTZ, Ph.D.**, *Department of Psychology, Harvard Medical School*
THE EMOTIONS
PERSONALITY
- MAURICE W. SCHWARTZ, Ch.E.**, *Consulting chemical engineer*
PLASTICS
- FRANCIS P. SHEPARD, Ph.D.**, *Professor Emeritus of Oceanography, Scripps Institution of Oceanography*
THE DEPTHS OF THE SEA
SEASHORES
- DENNIS SIMANAITIS, Engineering Editor, Road & Track
THE AUTOMOBILE**
- FERDINAND L. SINGER, Science writer
MOTION**
- BERNHARDT G. A. SKROTZKI, M.E.**, *Former Engineering Editor and Managing Editor, Power*
ELECTRICAL ENERGY
- W. H. SLABAUGH, Ph.D.**, *Professor of Chemistry and Associate Dean, Graduate School, Oregon State University*
METALS AND NONMETALS
- HAROLD T. U. SMITH, Ph.D.**, *Professor of Geology, University of Massachusetts*
HOW WIND CHANGES THE LAND
- HILTON A. SMITH, Ph.D.**, *Dean of the Graduate School and Coordinator of Research, University of Tennessee*
CHEMICAL REACTIONS
- MURRAY SPIEGEL, Ph.D.**, *Professor of Mathematics and Chairman, Department of Mathematics, Rensselaer Polytechnic Institute*
CALCULUS
- R. CLAY SPROWLS, Ph.D.**, *Professor of Information Systems, UCLA*
COMPUTER PROGRAMMING
COMPUTERS
- ANTHONY STANDEN, M.S.**, *Executive Editor, Kirk-Othmer Encyclopedia of Chemical Technology*
ACIDS, BASES, AND SALTS
CATALYSIS
- HARLAN T. STETSON, Ph.D.**, *Former Director of Laboratory for Cosmic Research, Needham, Massachusetts*
THE ATMOSPHERE
- JAMES STOKLEY, D.Sc.**, *Associate Professor of Journalism, College of Communication Arts, Michigan State University*
ELECTROMAGNETIC RADIATION
LIGHTNING AND THUNDER
- IVAN RAY TANNEHILL, D.Sc.**, *Former Chief of the Division of Synoptic Reports of the United States Weather Bureau*
WIND
- SANFORD S. TEPFER, Ph.D.**, *Associate Professor, Department of Biology, University of Oregon*
FRUITS AND SEEDS
- JENNY TESAR, Free-lance science writer
FLOWERS**
- JAMES TREFIL, Ph.D.**, *Professor of Physics, University of Virginia*
QUARKS
- RICHARD G. VAN GELDER, Ph.D.**, *Chairman and Curator, Department of Mammalogy, The American Museum of Natural History*
ENDANGERED SPECIES
- WERNHER VON BRAUN**, *Rocket expert and pioneer in space research; former Corporate Vice President for Engineering and Development, Fairchild Industries*
SPACE STATIONS
- DON WATERS**, *American Radio Relay League, Inc.*
AMATEUR AND CB RADIO
- RICHARD WENDT**, *New York Times*
PRINTING
- FRED L. WHIPPLE, Ph.D.**, *Former Professor of Astronomy, Harvard University; Former Director, Astrophysical Observatory, Smithsonian Institution*
THE SOLAR SYSTEM
- ORAN R. WHITE, Ph.D.**, *High Altitude Observatory, National Center for Atmospheric Research*
THE SUN
- J. TUZO WILSON, Sc.D., F.R.S.**, *Director, Ontario Science Centre*
PLATE TECTONICS
- VOLNEY C. WILSON, Ph.D.**, *Former physicist, Research Laboratory, General Electric Company*
COSMIC RAYS
- HARRY J. WOLF, E.M., M.S.**, *Former mining and consulting engineer*
METALLURGY
- RICHARD S. YOUNG, Ph.D.**, *Chief Environmental Biologist, NASA Ames Research Center, Moffett Field, California*
LIFE ON OTHER WORLDS
- RICHARD G. ZWEIFEL, Ph.D.**, *Curator, Department of Amphibians and Reptiles, American Museum of Natural History*
AMPHIBIANS

GUIDE TO THE NEW BOOK OF POPULAR SCIENCE

Astronauts blast to the moon and bring back samples from another world; one person's heart beats in another's body; newspapers report work on creating new forms of life in a test tube... These are just some of the most startling of the advances in science that have excited people in the last few decades. Is nature now under human control? The first rumblings of an earthquake, the appearance of a new island where only ocean was visible, the discovery of a hitherto unknown form of life deep in the dark and unfavorable ocean abysses answer that question, once again revealing the power and mystery of nature.

The 12 sections of which THE NEW BOOK OF POPULAR SCIENCE consists present the major fields of science and discuss their applications in the world of today. The section Astronomy & Space Science, for example, explores people's long fascination with the heavens, explains what we have learned through the centuries, and discusses how this knowledge and the quest for more has led to the exciting age of manned space exploration. The section on Earth Science and the closely related Energy and Environmental Sciences sections take us on a tour of our home base, probe how its vast energy is being and can be used, and describe how nature and people both work to change the earth.

The arrangement of articles in each section provides, as far as possible, a logical, step-by-step presentation of the subject matter in a particular field. Each article, however, constitutes a unit in itself and can be read as such. The articles are essay length, providing a well-rounded introduction to a particular topic. Subheads and sideheads within the text of each article help give the reader a quick and concise overall view of the article's contents. The metric system of measurement, long used by scientists throughout the world, is used throughout THE NEW BOOK OF POPULAR SCIENCE.

Illustrations are a very important part of THE NEW BOOK OF POPULAR SCIENCE. Plentiful and beautiful color photographs and artwork as well as black-and-white illustrations make THE NEW BOOK OF POPULAR SCIENCE a very attractive set that invites you to wander through its pages. A description of a solar flare is made real when you see the striking yellow and red outbursts from the solar disk. The entire story of the exploration of the moon comes alive as you see the astronauts on the moon, gathering and examining lunar rocks. Complex concepts in chemistry and physics become simple and the human body reveals its marvelous organization through clear diagrams. Electron micrographs of viruses and bacteria, underwater photographs of deep ocean life, views of animals in their natural habitats provide a beautiful panorama of life's diverse forms. Throughout, the illustrations complement the text, explaining, expanding and beautifying it.

For readers who wish to have additional information on a given subject, the editors have provided a bibliography, called Selected Readings, at the end of each volume. It contains not only a list of informative books but also a brief evaluation of each book that is listed so that the reader may have some idea of what it contains.

An alphabetical index for the six volumes of THE NEW BOOK OF POPULAR SCIENCE is given at the end of Volume 6. It enables the reader to find easily and quickly specific items of factual information

in the articles. Instructions for the use of the index are given on its first page. The index makes it possible for the reader to obtain the fullest possible benefit from the set.

In the pages that follow, we list the 12 sections of THE NEW BOOK OF POPULAR SCIENCE, and we give a brief account of the contents of each.

ASTRONOMY & SPACE SCIENCE

Vol. 1

No area of science has excited the imagination of civilized earthlings as much as have astronomy and space science. We dream of dodging among stars and through intergalactic systems, of shifting into hyperspace, and of finding little, green, one-eyed creatures at the edge of the universe. At the same time, earth-bound events such as the sight of a clear, full moon, or a spectacular orange-and-purple sunset still bring out feelings of romance and beauty.

In the section Astronomy & Space Science, we look at our universe and what we know about it. We consider first the face of the sky and the ways we have to study the skies. Then we discuss the star that we call the sun and the celestial bodies that form the solar system. Turning then to other stars, we learn about their composition, their brightness, their movements in space. Then we explore some of the most mysterious events in space—phenomena such as black holes and quasars. Finally we consider how astronauts have taken their first steps into space. We look at rockets, at space probes, at the problems and successes of manned space exploration, and review what we have learned from our visits to the moon. We end this section wondering about the possibility of life on other worlds and about unidentified flying objects.

COMPUTERS & MATHEMATICS

Vol. 1

Our everyday lives are touched in many ways—sometimes without our realizing it—by mathematics and computers. We may tend to think of mathematics as an esoteric field of science. Yet we also use mathematics when we do such simple things as count change or measure recipe ingredients or share a restaurant check. We use calculators to do mathematics homework or balance checkbooks. We depend on computers to provide us with up-to-the-minute weather reports, to keep grocery shelves stocked, to keep our home fuel tanks filled. Yet often the word computer brings images of future robot takeovers and of a dehumanized, push-button society. In the section on Computers & Mathematics we learn some of the basic facts of these sciences and so learning dispel many of our fears. We explore some of the major fields of mathematics: arithmetic, algebra, plane and solid geometry, trigonometry, analytic geometry, and calculus. We discuss some of the applications of mathematics in statistics, probability, and game theory. Then we go on to learn about binary numerals and the tool based on the binary numeral system—the computer. We learn how to program a computer and how to use it.

EARTH SCIENCES

Vol. 2

The earth is our home and we are, of course, interested in its foundations, its landscape, its features, its crust. In the section on Earth Science, we examine the earth's crust and some of the ways it has formed and is still changing through the action of earthquakes, volcanoes, and other agents of the tremendous power within the

earth's interior. We look at some crustal features—mountains and caves, for example—and piece together theories on how they have formed and how they are now working to affect the landscape. Then we turn our attention to the protective envelope that surrounds the earth—the atmosphere. We study lightning, wind, rain, and other atmospheric occurrences before we begin our exploration of the earth's third major part: water. Hidden in the ground, coursing through rivers, locked in huge oceans with their own life cycles, water covers three fourths of the earth's surface. We explore its characteristics, its actions, its life. We close this section with a brief look at the early history of our home in space.

ENERGY

Vol. 2

Perhaps the greatest challenge facing us today is the quest for sources of energy to power civilization and at the same time preserve the earth for future generations. We open the section on Energy by taking an overall look at how much and in what ways we use energy, at how we now obtain it, and at what our future needs are likely to be. Then we explore the major conventional sources of energy: oil, coal, and natural gas. Then we turn our attention to some of the more exotic sources: geothermal energy and solar energy—heat from the earth's interior and directly from the sun. We also discuss nuclear energy—both fission and fusion. Is it safe? Will it answer future needs for energy? We end the section with a discussion of the way in which most of our energy is used—in producing electricity.

ENVIRONMENTAL SCIENCES

Vol. 2

Will earth be a suitable home for future generations? In the section on Environmental Sciences we examine our total environment, or surroundings, and see how the forces of nature and the hand of man have changed and are continuing to change the earth. First, we examine earth's natural resources—water, wood, minerals—and discuss how they can be both used and preserved. Then we see how nature's forces—running water and wind, for example—are carving the land, changing its face. We next see how human beings—through industry, agriculture, transportation, and other activities—are changing the earth. We wonder what will be left, and in what condition, for future generations. We take some plant and animal species as examples illustrating the role of a changing environment and discuss how the existence of some species is threatened and the steps that can be taken to safeguard them and other endangered species. We also explore some of the areas now set aside as specially protected regions where nature wins and people are allowed little or no influence—so-called wilderness areas.

PHYSICAL SCIENCES

Vol. 3

What do all things—living and nonliving alike—have in common? They are all forms of matter—that which occupies space. The different forms of matter possess and can be made to possess energy. Matter and energy are the twin bases of all things. In the Physical Sciences section we study chemistry, "the matter science," and physics, "the energy science." We explore matter's different forms and its basic component—the atom. We see how various elements combine to form acids, bases, salts, solutions, and colloids.

As we turn more particularly to the study of energy in the physics subsection, we see how energy is the basis of work and movement. We investigate the nature of electricity and its relationship to magnetism. We explore how energy is involved in such phenomena as heat, sound, optics, and color. Then we turn to what is probably the scientific theory most closely associated with modern science—relativity—and try to understand its basic premises and applications. We end the section exploring two of matter and energy's most intriguing phenomena: the world of plasma, or superhot, energized gases; and the world of supercooled, superconducting elements.

BIOLOGY

Vol. 3

What is life? How did it arise and diversify into the countless forms it now takes on land, in water, and even in the air? In the General Biology section we explore the nature of life. We study its simplest and most basic unit—the cell—learning its structure, tracing the steps of its activities, and marveling at its complexity. We consider how living things—small and large alike—develop from a single cell and how the many varied organisms have attained their present organization through a long series of changes to which we give the name evolution. We see how organisms react to their environments and to one another in a particular community and habitat, forming a complex and completely interdependent web of life. Finally, we see how life forms are classified, or grouped.

PLANT LIFE

Vol. 4

The plant kingdom includes forms from the tiny diatom visible only through a microscope to the mighty redwood trees towering over one hundred meters high. In the section on Plant Life we see how these two seemingly diverse forms are related—how they are both plants—and go on to study the characteristics of all plant life. We start with the soil, the anchor for all land-dwelling plants, and proceed from the simplest plants to the highest forms. We see how algae and fungi form an essential part of the cycle of life on earth and how ferns, which beautify the earth in dense and luxurious growths, begin to show some of the characteristics of higher plants. We devote particular attention to the seed plants, the dominant form of vegetation on earth. We study their roots, stems, leaves, flowers, and fruits. We also consider how various plants adapt to their environments. We end this section with a discussion of some plants of particular importance and interest to humans—vegetables, cacti, houseplants, and trees.

ANIMAL LIFE

Vol. 4

The animal kingdom is vast and includes strikingly diverse forms—from tiny one-celled protozoans to highly complex mammals—including people. The section on Animal Life surveys the vast panorama of animal life. The story begins with the invertebrates, or animals without a backbone: protozoans, sponges, starfish, worms, mollusks, lobsters and other crustaceans, spiders, and many types of insects. Then comes the backboneed animals, or vertebrates: fishes, amphibians, snakes, and birds—similar in some ways, yet each exhibiting a unique adaptation to a particular way of life.

In the section on Mammals we explore in somewhat greater detail many of the most important mammal groups. We start with two highly unusual types of mammals—egg-layers and marsupials. Then we discuss the dominant and highly varied placentals—rabbits, rodents, aquatic mammals, dogs, bears, weasels, cats large and small, hoofed animals—to name just a few. We end the Mammals section with a discussion of the primates, the mammal group to which we belong.

HUMAN SCIENCES

Vol. 5

People are similar in many ways to animals and are in fact a part of the primate mammal group. Yet people are different from even their nearest mammal relatives. In the section on Human Sciences we deal with peoples' similarities to other animals and also with their differences. We start with a survey of the human body, studying its structure and how the various organ systems work. We then consider what personality is, what emotions are, and how memory works. We go on to see the human life cycle—heredity, development, growth, decline, and old age. We find that many factors—diet, sleep, and exercise—play important roles in our health. We see how a person's mind and body interact and how psychological stress can affect physical health. We discuss particular problems such as alcoholism, smoking, and drug abuse. We also consider a few of the particular diseases we commonly encounter—allergies, influenza, cancer, and arthritis—and end with a brief discussion of organ transplants and antibiotics, just two of the many advances of modern medicine.

TECHNOLOGY

Vol. 6

It is through advances in technology that most of us are made aware of the achievements of science. Technology affects our everyday lives, taking the discoveries of science and putting them into our homes, offices, factories, and cars. The Technology section begins with a discussion of some of the major divisions and functions of modern engineering. Then we survey some of the oldest applications of science—the early technologies in which people found materials for their homes and built homes and other buildings. Next we explore the field of transportation, tracing the development of the automobile, railroad, airplane, and other means of travel. Turning our attention to communications, we see how the world of electronics has provided us with a wide range of communication devices.

Continuing our look at applied science, we see how the chemical industry provides us with a wide range of products, how the use of plastics has changed many aspects of our daily lives, how the use of ultrasonics, fiber optics, and liquid crystals is affecting and will continue to affect our ways of doing things. We see how the development of X rays revolutionized medicine and other fields of science and how radiocarbon dating techniques provided us with means of dating and piecing together the earth's early history. Finally we turn our attention to cybernetics and end by discussing the possibilities of automatic self-regulating systems.

Volume 1

Contents

| | |
|---------------------------|-----------|
| Astronomy & Space Science | 1 - 329 |
| Computers & Mathematics | 330 - 483 |



ASTRONOMY & SPACE SCIENCE



both photos, NASA

The planet Mars began to reveal its secrets as the U.S. Viking space probes approached and landed on the planet. The photomosaic at left reveals Vallis Marineris, the "Grand Canyon" of Mars (upper left of darkened area), and the large impact basin Argyre (lower left of dark area). The photo above shows a Martian sunset.

| | |
|---------|-----------------------------|
| 2-12 | The Study of the Universe |
| 13-16 | Origin of the Universe |
| 17-24 | The Night Sky |
| 25-35 | The Constellations |
| 36-44 | Telescopes |
| 45-53 | Observatories |
| 54-57 | Planetariums |
| 58-64 | Radio Astronomy |
| 65-78 | The Sun |
| 79-81 | Copernicus |
| 82-89 | The Solar System |
| 90-93 | Mercury |
| 94-96 | Venus |
| 97-105 | Earth |
| 106-116 | The Moon |
| 117-122 | Mars |
| 123-128 | Jupiter |
| 129-136 | Saturn |
| 137-141 | The Outermost Planets |
| 142-148 | Eclipses |
| 149-157 | Comets |
| 158-160 | Asteroids |
| 161-167 | Meteorites and Meteors |
| 168-173 | The Calendar |
| 174-178 | X-Ray Astronomy |
| 179-192 | The Stars |
| 193-199 | Variable Stars |
| 200-205 | Interstellar Space |
| 206-214 | The Milky Way |
| 215-222 | Galaxies |
| 223-229 | Cosmic Rays |
| 230-235 | Pulsars |
| 236-239 | Quasars |
| 240-245 | Black Holes |
| 246-253 | Rockets |
| 254-260 | Guidance in Space |
| 261-268 | Space Satellites and Probes |
| 269-272 | Communications Satellites |
| 273-284 | Manned Space Flight |
| 285-293 | Space Stations |
| 294-303 | Exploration of the Moon |
| 304-315 | Geology of the Moon |
| 316-325 | Life on Other Worlds |
| 326-329 | Unidentified Flying Objects |



California Institute of Technology and Carnegie Institution of Washington

Some of the most interesting and beautiful objects of the universe are glowing masses of dust and gas known as nebulae. Here are two nebulae in the constellation Cygnus. Above: the Veil Nebula; at right: the North American Nebula.

THE STUDY OF THE UNIVERSE

by Katherine Haramundanis

When you look up at the sky on a clear night, what do you see? Most of the dark night sky is dotted with twinkling points of light, which we call stars. The moon may be in the sky as well—perhaps as a thin arc, perhaps as a full circle of light.

The brightest point of light might be the planet Venus, and a reddish one would be the planet Mars. Except for the other planets, all the other points of light are stars.

Sometimes you can see a faint, glowing band across the overhead sky when it is clear and dark enough. It is made up of stars too far away or too small to be seen as separate points of light. This band is known as the Milky Way.

What is it like out there? What are the stars? How far away are they? How did they come to be? Will they go on shining forever? And what of space itself? Does it

come to an end somewhere? Or does it go on without end?

OBSERVING THE SKY

Try to imagine, for yourself, how early stargazers learned about what they saw in the sky.

Perhaps you would start simply by observing the night sky over a period of time. After a while, you would begin to notice that the brighter stars seemed to form patterns in the sky. You would remember these patterns when you saw them again. That is what early stargazers did, anyway. They traced out star patterns and named them after gods and heroes, or animals and familiar objects. They called these star patterns *constellations* (from the Latin, *com stella*).

The names given to constellations in ancient times are still in use today—Scor-



Société astronomique de Suisse

pius, Leo, Orion, and 85 others. Most constellations do not look like the mythical figures they are named after—at least not to our eyes. However, they are useful even now for learning our way around the night sky.

Having learned to pick out certain stars by their positions in constellations, you would be able to keep track of their movement across the sky from night to night. This is what the early stargazers did. They also kept track of the movements of the sun and the moon. The star watchers also kept records of all these heavenly bodies over periods of many years.

In this way they came to realize that events in the sky repeat themselves over and over again. From repeated observations about the times when the sun, moon, and planets appear, people slowly came to develop the first calendars for keeping track of time and the seasons. In fact, that was one important reason why people studied the sky.

EARLY IDEAS OF THE UNIVERSE

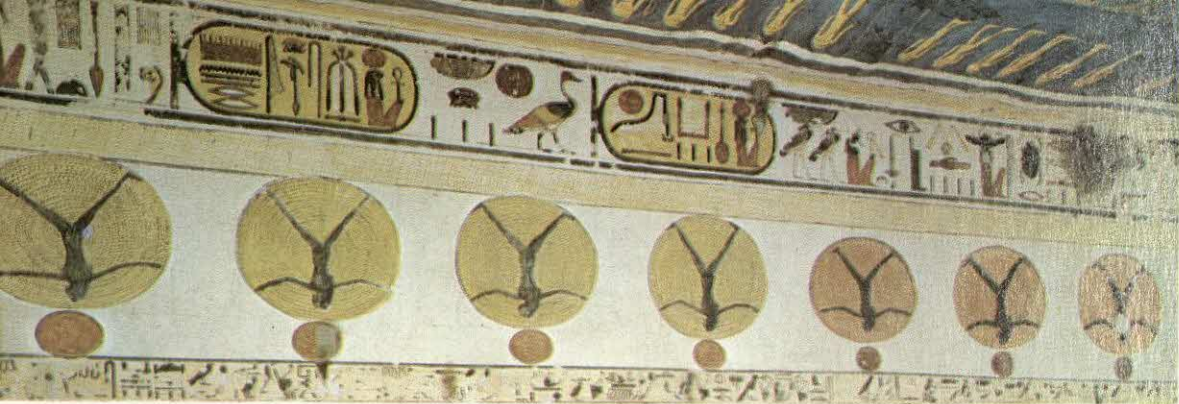
The first civilization to gain a real understanding of the sky's objects and their motions was that of the ancient Greeks. Some of the Greek astronomers were very careful observers, and kept long and de-

tailed records. But most of the Greek thinkers were mainly interested in developing theories that could explain the universe.

Some of their theories came close to what astronomers now believe to be true. The Greeks did this without using any of the instruments we associate with astronomy today, such as the telescope. For example, most Greek thinkers came to realize that the earth is actually spherical, not flat. One of their astronomers even made a fairly exact calculation of the size of the earth by using the methods of geometry.

Another common belief of ancient peoples, including the Greeks, was that the earth is the center of the universe. After all, that is the way it really looks. The heavenly bodies appear to be circling the earth. The planets look like points of light moving slowly among the stars from night to night. They also seem to be circling around the earth. If you knew nothing about astronomy, it would be only natural for you to think—as the ancients did—that the earth is the center of everything—the universe.

Some Greek astronomers, however, thought of the universe in a different way. In the fifth century B.C., the Greek astronomer Anaxagoras decided that the sun and moon and planets were not simply lights



Lessing: Magnum Photos

Astronomy in ancient Egypt. Painting on the tomb of Ramses IX, an Egyptian pharaoh, illustrating the death and resurrection of the sun.

in the sky. Instead, he described them as solid bodies like the earth.

One century later, the Greek thinker Heraclides offered another new idea. He suggested that the earth does not stand still. Instead, wrote Heraclides, the earth rotates—that is, it spins like a top. He also suggested that the planets revolve around the sun. However, he still thought that the sun, with its family of planets, revolved around the earth.

Aristarchus, a few years later, went much further toward a modern view of the universe. He stated that the earth was not the center of all things. Instead, he thought that the earth revolved around the sun.

THE PTOLEMAIC SYSTEM

The idea suggested by Aristarchus, that the earth actually moves through space around the sun, was a very strange one for people of his time. They could see the sun move across the sky. Most of them were not ready or willing to accept a new theory. The common belief continued to be that the earth stands still, and Aristarchus' theory was ignored.

In the second century A.D., a Greek scientist named Ptolemy brought together all of the astronomical and other scientific knowledge of his time in a series of books. The picture of the universe that he developed came to be known as the *Ptolemaic system*. It became the accepted view of the universe for about thirteen hundred years.

In the Ptolemaic system the moon, sun, stars, and planets were thought to travel in perfectly circular paths around the earth. There was no scientific reason for

this notion. It simply fitted in with the Greek idea that a circle was a perfect shape. Today we know that the planets travel in *elliptical* paths—like somewhat flattened circles—as they move about the sun. But some Greek astronomers were aware that the planets' paths, or orbits, did not really seem to be perfect circles around the earth. Therefore the believers of the Ptolemaic system tried to figure out some way to explain this fact.

There was something else that seemed “wrong” about the way some planets moved in the sky. Sometimes they did not appear to be revolving around the earth at all, but moving backwards.

To account for noncircular orbits and for the peculiar movements of some planets such as Mars, the Ptolemaic system used the idea of an *epicycle*. In this, a planet moved in a circle. The center of this circle moved along another circle. You can see how this works, by looking at the drawing on page 81.

This notion of epicycles, used in the Ptolemaic system, was a complicated one. There was no scientific reason behind it. Its only purpose was to make it possible for people to describe the paths of the planets as a system of perfect circles.

Today we know that the planets do not revolve around the earth. We also know that planets farther from the sun have larger orbits and move more slowly than planets nearer the sun. Thus as Earth and Mars—a more distant planet—both move around the sun, Earth keeps catching up with and passing Mars relative to the background of stars. When it does so, Mars

seems to stop moving against the background of stars. Then Mars seems to move backward for a while, pause, and start moving forward again.

THE COPERNICAN REVOLUTION

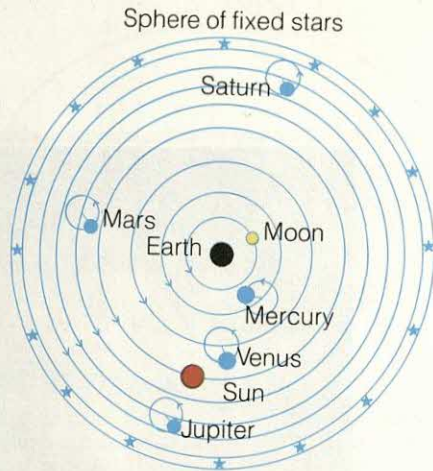
No real advances were made in the science of astronomy during the period we know as the Middle Ages. However, among the changes that marked the close of the Middle Ages was a growing scientific interest in the natural world. In time, the Ptolemaic system itself came to be challenged. The most famous name involved in this challenge is that of the 15th-century Polish astronomer, Nicholas Copernicus.

What Copernicus did was to state that the sun, not the earth, is the center of the universe. He said that the earth moves through space around the sun. He also said that the earth spins, or rotates, like a top. As mentioned above, these ideas had already been suggested in ancient Greece. Copernicus, however, worked out the ideas in detail. He did this before the invention of the telescope and other instruments of modern astronomy.

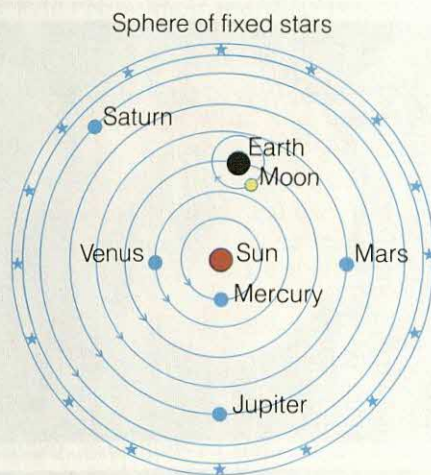
Ptolemy had argued that the earth could not be rotating. Otherwise, clouds could not move eastward, because the air would be blowing in the other direction all the time. Copernicus pointed out that the clouds and all the rest of the atmosphere were really part of the earth and were rotating along with it.

Copernicus also showed that the apparent motion of Mars in the sky could be much more easily explained. This he did by stating that the earth and Mars travel around the sun, but at different speeds. In this way the apparent loops, or epicycles, that Mars made in the sky simply became an optical illusion.

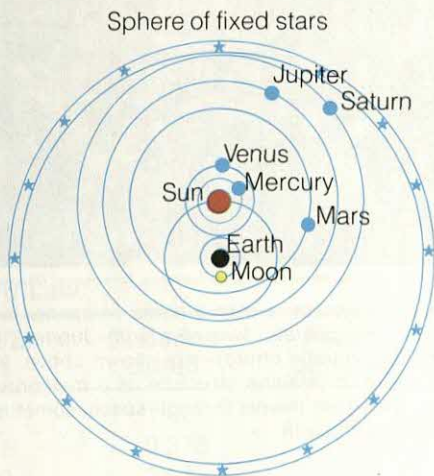
Copernicus did not give up one Ptolemaic idea, however. He, too, could not believe that planets did not move in perfect circles. Because of this, he was not able to work out, mathematically, the orbit of a planet such as Mars. Copernicus did not publish his work until the year of his death, although fellow scientists were already aware of his views. The book, *De revolu-*



Ptolemy thought that the earth was at the center of the universe and that the sun and planets revolved around the earth.



Copernicus demonstrated that the earth, moon, and planets revolved around the sun. He was wrong, however, in thinking that the orbits were perfect circles.



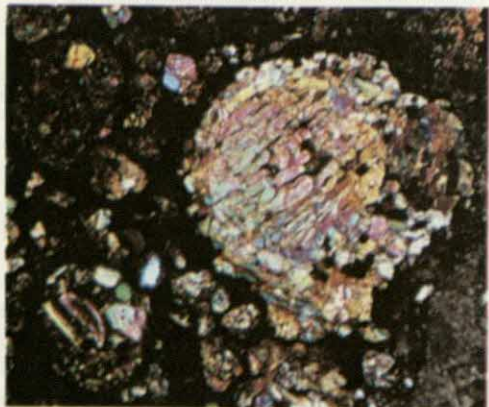
The astronomer Tycho Brahe combined the systems of Ptolemy and Copernicus. His universe was earth-centered.



© California Institute of Technology & Carnegie Institution of Washington



© California Institute of Technology & Carnegie Institution of Washington



John Wood, Smithsonian Institution Astrophysical Observatory

The solar system is man's home in space. It includes nine planets, two of which—Jupiter (top) and Mars (middle photo)—are shown above. Bottom photo: crystalline structure of a meteorite, a small body that travels through space, sometimes colliding with earth.

tionibus orbium coelestium ("On the Revolutions of the Heavenly Spheres"), may be said to mark the beginning of modern astronomy.

Copernicus' idea of a sun-centered universe was not accepted right away by most astronomers and the rest of society. Tycho Brahe, for example, was an important observational astronomer who lived in the last half of the 16th century. He refused to believe that the earth was not the center of the universe. He suggested, instead, that the other planets revolved around the sun, and that the sun in turn revolved around the earth.

KEPLER, GALILEO, AND NEWTON

An assistant of Brahe named Johannes Kepler, however, accepted the Copernican theory. Being a mathematician, Kepler worked out three laws dealing with the orbits of planets.

By carefully studying Brahe's numerous observations of Mars, he came to realize that it moves in an elliptical path, rather than in a circular one. Kepler's first law of planetary motion states that a planet moves in an ellipse with the sun at one focus.

Kepler's second law describes the varying speed of a planet in its orbit. When a planet is nearest the sun, it is moving the fastest. And when it is farthest from the sun, it is moving the slowest. (Study the diagram and its caption.) Kepler was coming close to the idea of gravitation without explicitly announcing it.

Kepler's third law relates a planet's distance from the sun with its *period*. A planet's period is the time it takes to make one revolution around the sun. The law states that the period squared (P^2) equals the distance cubed (D^3). Examples of this law are given in the table on the next page, which uses the distance of the earth to the sun as one a.u. (astronomical unit) and its period as one year.

While Kepler was developing these laws, the Italian scientist, Galileo, was making another great contribution to astronomy. He used the recently invented telescope to look at the planets, the moon,

and the sun. (The last was a bad mistake. He looked at the sun directly with the telescope and blinded himself for a while.) What he saw in the sky did not agree with the teachings of the ancient Greeks. For example, he saw that the moon is not a perfect sphere at all. Instead, he observed that it had valleys and mountains, like the earth.

Galileo also discovered that Venus goes through phases, like the moon does. He observed the moons of Jupiter for the first time. And with his telescope he learned that there are many more stars in the sky than the naked eye can see.

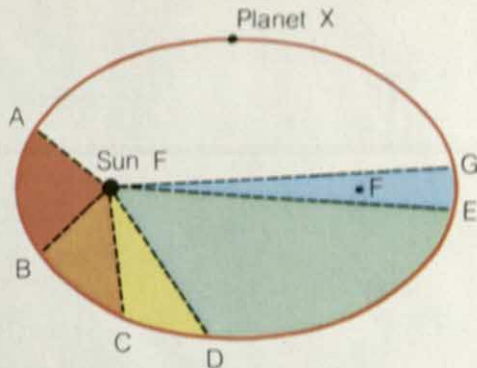
In the late 17th century a great English scientist, Isaac Newton, developed the law of gravitation. This law states that all objects are attracted to all other objects by the force of gravitation. The strength of this force depends on how much matter the objects contain and on the distances among them. The law explains why planetary and lunar orbits are elliptical. It explains the motions of all objects in the universe.

This scientific explanation of the motions of the earth, sun, moon, and planets lasted without change until the 20th century. By the time of Newton's death, he had become one of the most honored of all scientists.

TELESCOPES AND CLOCKS

Modern astronomy grew out of the work of such scientists as Copernicus, Kepler, Galileo, and Newton. Its development also depended on two important inventions: the telescope and an accurate clock.

The invention of the telescope opened up the skies to astronomers. The first telescopes were simple two-lens systems. One lens focused the light. The other, some distance away, was an eyepiece, which magnified the focused image. Such telescopes are called *refractors*.



According to Kepler's second law, a planet, X, sweeps out equal areas (shaded regions) in equal time intervals. It is apparent that the planet's speed must be greater between A and B than between E and G.

No longer were the planets mere points of light. Now they and their moons could be seen as other worlds—solid-looking bodies like the earth. As telescopes improved, more and different sky wonders were revealed.

Newton himself made an important contribution to the design of telescopes. He developed a telescope that uses a mirror to reflect and focus light. Such telescopes are called *reflectors*. Today the largest optical telescopes are reflectors.

For a long time, however, refractors continued to be the most common kind of telescope, because people did not know how to make large and accurate mirrors. Instead, astronomers tried to develop better lens systems for their refracting telescopes.

Toward the end of the 17th century, the Dutch astronomer Christian Huygens invented the first pendulum clock. Until that time, it had not been possible to measure time precisely. Such an ability is very important in modern astronomy.

As in the other exact sciences, measurements in astronomy have to be precise to be useful. Scientists have to keep accurate track of when events take place and how long they last. Thus with these two instruments—the clock and the telescope—

Kepler's Third Law

| Planet | Distance from Sun (a.u.) | Period in years | Distance cubed | Period squared |
|---------|--------------------------|-----------------|----------------|----------------|
| Earth | 1.00 | 1.00 | 1.00 | 1.00 |
| Venus | 0.723 | 0.615 | 0.378 | 0.378 |
| Mars | 1.524 | 1.881 | 3.54 | 3.54 |
| Jupiter | 5.20 | 11.86 | 140.66 | 140.66 |



S.A.S.

The universe consists of perhaps 10,000,000,000 star systems, or galaxies. The solar system is in the galaxy called the Milky Way. Above is the image of a spiral galaxy.

the real growth of modern astronomy was under way.

WHAT OF THE STARS?

Through the early years of this growth, however, the picture of the universe held by astronomers was still quite different from the one we have today. That is, the nature of the starry sphere was still unknown. Even by the time of Newton, no one understood what the stars really were. Figuring this out led to the next great change in our view of the universe.

In ancient myths, the stars were sometimes described as dots painted on the sphere of the sky. Sometimes they were said to be holes in the sky, letting in light from a great fire beyond. No one knew how to figure out the distances to the stars. In fact, not until the time of Newton had people figured out the actual distance to any heavenly body except the moon.

Newton was the first to give a scientific basis to the idea that stars are suns, like our own sun, and that they are very far away. His friend, the astronomer Edmund Halley, then discovered that the stars move in relation to each other's position in the

sky. He did this by studying old star charts and finding out that the positions of some stars had changed over a period of many years. The celestial sphere could no longer be thought of as going on forever without change. The universe was now seen as an unending stretch of space, through which the stars were moving in different directions.

Astronomers probed the universe of stars with more powerful telescopes through the next two centuries. By the late 18th century an English astronomer, William Herschel, described the known universe as being shaped like a "millstone." That is, he found that the stars seemed to be scattered through space around the sun in the form of a thick disk. This disk contained all the individual stars we can see, plus the more distant stars that make up the Milky Way.

So far as their instruments could tell them, however, astronomers still thought that the sun was at the center of the universe. At any rate, stars seemed to thin out in every direction away from the sun. Thus the sun appeared to be at the center of this gathering of stars.

DISTANCES TO THE STARS

But how far away, actually, were all these stars? All that scientists such as Copernicus could say was that they must be far away. They were able to say this for a simple reason. The positions of the stars and the distances among them did not seem to change at all from one time of year to another.

What this tells us can be understood if you imagine yourself in a moving car. When you look at the horizon, you notice that objects nearer to you change their position against the background of more distant objects. In the same way, you are traveling through space as the earth moves around the sun. If there were stars fairly close by, the nearest ones would appear to shift their position against the background of more distant stars as you observed them throughout the year. This apparent change of position with background stars is called *parallax*.

By the 19th century, however, no telescopes had yet been able to detect any such shift for any star in the sky. This meant that even the nearest stars must be so far away that the earth's motion around the sun—the distance it travels—makes no difference in comparison.

The first actual measurement of the distance to a star was not accomplished until 1838. The work was done by an amateur German astronomer, Friedrich Wilhelm Bessel. He chose a star known as 61 Cygni. It is in the constellation Cygnus. This star had been observed to be changing its position in regard to other stars; that is, it showed a large proper motion. This probably meant that it was nearby—again, as stars go. (The proper motion of a star is the observed motion, over a period of time, relative to “fixed” stars.)

Bessel carefully observed 61 Cygni and its background stars as the earth progressed in its elliptical orbit around the sun. He achieved what he set out to do. He detected a slight shift in the star's position against the background of more distant stars. Simple geometrical methods then told him how far away 61 Cygni must be. The star was 11 light-years away.

(Astronomers use light-years to describe star distances. One light-year is the distance light travels in one year—about 9,500,000,000,000 kilometers.)

NEW GAINS IN ASTRONOMY

In the 18th and 19th centuries, the picture that people had of the universe began to be more like the one we have today. Great observational astronomers, such as William Herschel, discovered that many of the stars we see are actually two-star systems, or *double stars*. Three-star and even more complicated systems were also discovered. Herschel also observed several spherical groups of stars in which the stars were packed closely together. We know these now as *star clusters*.

Herschel also observed the huge clouds of dark or glowing dust and gas that we call *nebulae*. Finally, he discovered another planet in our solar system, which he named Uranus. Still another planet, Neptune, was discovered a few decades later, in 1846.

Probably no one astronomer or physicist can claim the development of the theory that the sun and stars are enormous masses of rotating gas, producing light and heat and other forms of energy.

In the early 19th century Herschel estimated such a large figure for the energy coming from the sun that the sun would be consumed in less than 5,000 years if the energy came from ordinary oxidation.

In 1802, Joseph von Fraunhofer observed spectral lines in solar spectra. In 1859, Gustav Kirchhoff stated a general law that connected spectral lines with the absorption and emission of light and emphasized that each atom had a unique spectrum. In 1861, Kirchhoff and Robert Bunsen made the first chemical analysis of the solar spectrum and laid the foundation for the use of the spectrum in astrophysics.

By 1864, John Herschel and Father Secchi suggested that the sun was totally gaseous. During the next decades of the 19th century, William Huggins revolutionized astronomy by using the spectroscope to study the chemical composition of many stars.

With the development of photography in the 19th century, astronomers gained a major new tool. By using cameras, they could take pictures of the sky and study them as long as they wanted. They no longer had to remain at the telescope for long hours and try to remember what they saw. Photographs gave them a permanent record. By exposing film to the sky for long periods of time, they could detect very faint sources of light. Such sources include dim, nearby stars as well as the most distant known objects in the universe.

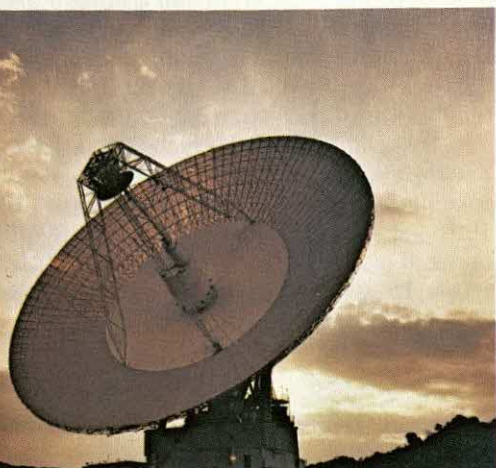
The 19th century also gave astronomers another tool for studying the sky—the spectroscope. A *spectroscope* is an instrument that breaks light up into the colors of the spectrum—red, yellow, blue, and so on. One end of the spectrum is red. The other end is violet. The other colors fall in between.

The spectra of the sun, stars, and other objects tell astronomers a great deal about these objects. The reason is that when elements are heated until they glow, they emit distinctive colors and dark or bright lines.

By using a spectroscope, scientists can identify the elements in a star. Thus, studying the spectrum of a star tells what a star is made of. Astronomers also use spectroscopes to learn about temperature and other conditions on a star. With the spectroscope, astronomers could begin to study the physics and chemistry of celestial objects as well as their visual appearance.

Radio telescopes are valuable instruments for probing the far reaches of the universe. Many discoveries, such as pulsars, were made with them.

NASA



Even with all these advances, our view of the universe by the end of the 19th century was still very different from the one we know today. People still thought of the sun and its planets as more or less at the center of the universe.

THE UNIVERSE WE KNOW TODAY

The 20th century brought a major change to our understanding of the universe. In 1917, an American astronomer, Harlow Shapley, gave us a new understanding about our star system, the Milky Way. His interpretation of many astronomical observations was that this star system, or *galaxy*, is not centered around our sun at all. Instead, our sun is about three-fourths of the way toward the outer edge of our galaxy. The last hint of the ancient view of an earth-centered—or, at least, sun-centered—universe was now removed.

There was more to come. In the 1920's, another American astronomer, Edwin Hubble, showed that our Milky Way, with its hundreds of millions of stars, is not the only star system in the universe. Rather, it is just one of many such galaxies. A very few of them have always been visible to the naked eye, but it took the powerful telescopes of the 20th century to show that they are actually made up of millions of individual stars.

The other galaxies have a wide range of shapes and sizes. Many are spiral-shaped, as our own was found to be. Others are elliptical, and others are spirals crossed by a broad band of stars at right angles to the main disk. The deeper the telescopes probed into space, the more galaxies they discovered. Astronomers now estimate that there might be 10,000,000,000 galaxies.

In 1926 the American astronomer V. M. Slipher discovered with a spectrometer that the light spectrum of distant galaxies is shifted toward the red end. This effect is now known as the *red shift*. Most astronomers interpret this effect to mean that galaxies are moving away from us at great speeds. The most distant galaxies produce a very large red shift, which must mean a very high speed of recession.

Among the more exotic objects in the

sky are the *pulsars*. They are thought to be rapidly rotating stars made up mostly of neutrons. The observed pulses appear to be produced by their rotations. Pulsars are members of a class of galactic objects called *variable stars*, because they change in brightness. One kind of variable star is the *nova*, which suddenly increases dramatically in brightness. We do not know why novas explode in brightness.

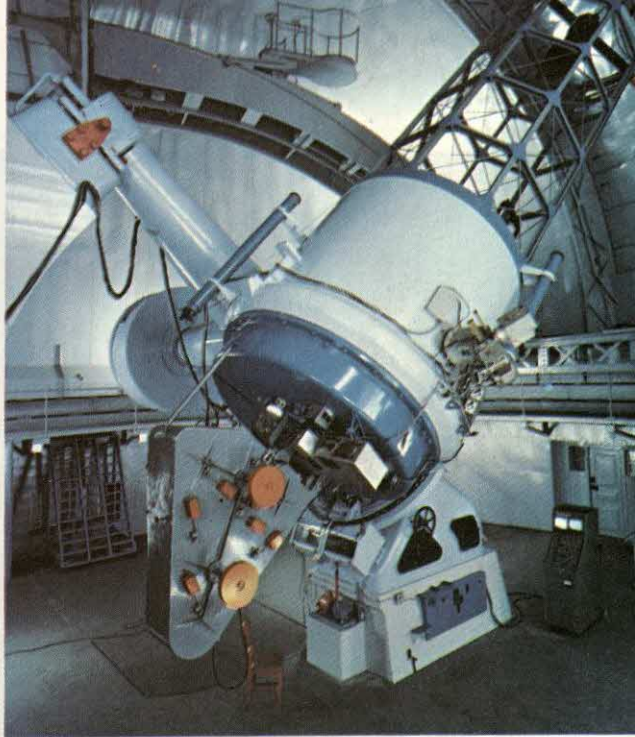
NEW OBSERVING METHODS

Scientists today have many different kinds of instruments with which to study the universe. No longer are astronomers limited to studying the light coming from celestial objects. Astronomers can now observe the radio waves, ultraviolet rays, infrared rays, X rays, and other radiations given off by the objects of the universe.

Many of the objects in our galaxy and in others have been observed in radio waves. *Quasars* were first observed by their radio waves, and later identified by light waves. They are remarkable because they appear to be moving at speeds approaching the speed of light. Quasars appear like stars on a photographic plate, but they also have a very large red shift. Some astronomers suggest that they are associated with galaxies. The name comes from "quasistellar object."

Astronomers are now detecting X rays from points in the universe that appear completely black. They now call such points *black holes* and theorize that they are burned-out stars. The mass of a black hole has been crushed into a small volume, but its gravitational force is so immense that it pulls back any light waves that might be leaving it. Astronomers also believe that a black hole's gravitational force sucks in galactic matter and spews it out as X rays.

Think back a bit. This has been quite a change, from the picture of the universe in ancient times to the one we know today. The earth used to be thought of as the center of all things, with the sun and planets and moon and starry sphere all circling around it. The earth today is seen as merely one of several planets revolving around an average-sized star in the outer regions of



David Dunlap—University of Toronto

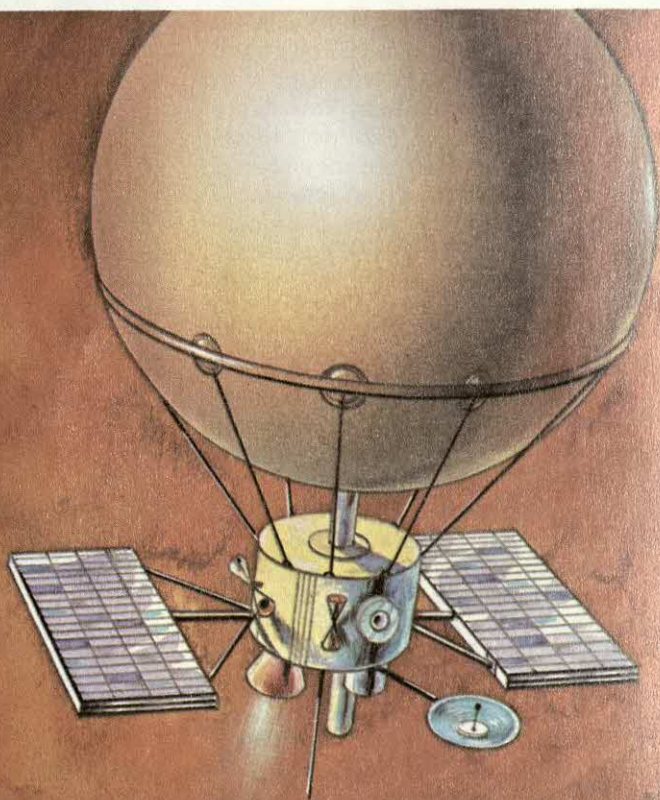
Telescopes today, such as this one at the University of Toronto, are guided automatically, and record images of the sky on photographic plates.

a fairly ordinary galaxy. Far from being at the center of all things, the earth might now be thought of as a mere speck of matter adrift in an enormous universe.

HOW LARGE AND HOW OLD?

Indeed, the size of the known universe is really too enormous for us to imagine. How large is it actually, in terms of light-years? Astronomers do not have the final answer to that question. They are not even certain of the methods they use to determine the very greatest distances. There are also several different theories about how large the universe can be. And this also applies to determining the age of the universe, which some astronomers believe to be about 16 billion years old.

Our own galaxy, the Milky Way, is about 100,000 light-years across at its widest part. Andromeda, the closest galaxy to ours, is 2.2 million light-years away. The most distant galaxies appear to be 10 billion light-years away. But with every telescopic improvement, the distance it "sees" increases. What could lie beyond?



An artist's concept of a balloon-type robot probe that would float over the surface of a celestial object being explored. Unmanned space probes are already playing an increasingly important part in our study of the universe.

Absolute nothingness?

People have been developing new answers to such questions in the course of the 20th century. Albert Einstein, in particular, developed the theory of relativity, upon which most modern views of the universe are based. There is no easy way to describe these views. And even these most modern theories are not all in agreement. Some say that the universe is a closed system—that it does not go on spreading outward forever. Others say that the universe is expanding from some early event and that it may never contract again. While a third group suggests that the universe is oscillating, now in an expansion phase, later to contract.

There is another great question for astronomers to solve. That is, where did all the stars and galaxies come from? Did they always exist? Or did the universe have a beginning, and will it come to an end? A number of major theories have been developed to answer these questions. You will be reading about them in the article that follows.

People are now sending instruments including telescopes into space to orbit the earth or to reach other planets. And new data keeps pouring in from all these sources. The information often answers old questions while raising many new ones. Our great human venture into the universe has really only begun.

ORIGIN OF THE UNIVERSE

by George B. Field

Did the universe originate suddenly in an enormous primeval explosion, thousands of millions of years ago? Or has the universe always been in process of creation, without a definite beginning or end? Exponents of the first idea, called the "big-bang theory," believe that all the matter in the cosmos once formed a compact mass, which some have likened to a huge "atom" of sorts. This mass then exploded, forming a vast fireball. In a few minutes, perhaps, matter was scattered across immense stretches of space. Today, the stars, galaxies, and planets that were formed from this material still have the motion that resulted from the explosion and are speeding away from each other at tremendous velocities. The different elements developed from the primitive matter that exploded.

On the other hand, followers of the continuous-creation, or "steady-state," theory, say that the universe has always been much the same for long ages, and that matter, namely hydrogen, is constantly being created, apparently from nothing. This material forms the stars and galaxies and arises more or less uniformly throughout the cosmos. In this article, we shall assess the scientific standing of the big-bang and

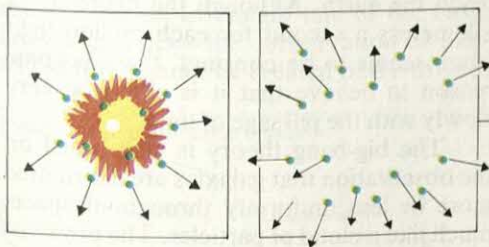
steady-state theories in relation to the latest astronomical research.

BIG-BANG THEORY

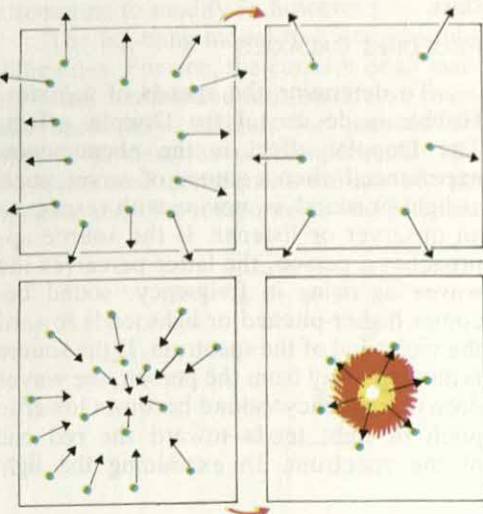
Theories of the universe form a discipline known as cosmology. Einstein was the first truly modern cosmologist. In 1915 he completed his general theory of relativity. It was then applied to the problem of the distribution of matter in space. In 1917 it was determined mathematically that there seemed to be a more or less uniform mass of material which is precariously balanced between the attractive force of gravitation and another, unfamiliar, force of cosmic repulsion, or pushing apart. In 1922, a Russian physicist came up with another solution to the problem, in which repulsion does not play a role; instead, there is an expanding universe where all particles are flying away from each other at high speeds. The expansion is constantly slowing because of the attractive force of gravity. Earlier, the particles had been moving outward even faster. In this model of the universe, the expansion started at a unique moment in the past—the so-called "big bang".

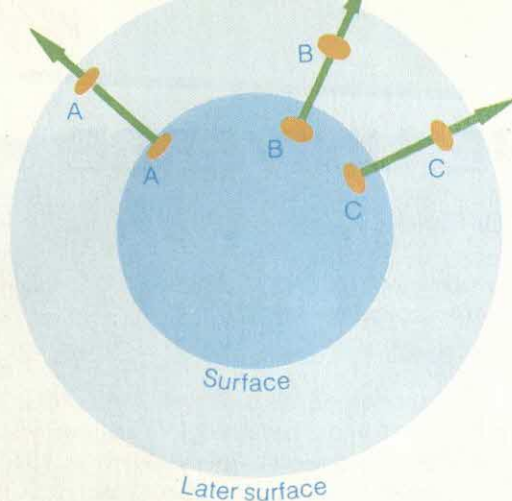
The big-bang theory seemed so contrary to the astronomical knowledge of the

Big Bang (two panels below, two right): the universe as we know it started with a big bang and keeps expanding forever.



Cyclic (two above, two right): universe follows endless cycle: big bang, expansion, contraction, big bang, expansion.





Expansion of the universe according to the big-bang theory is compared here to a balloon undergoing inflation. The specks A, B, and C represent galaxies at relatively fixed locations.

period, that at first it attracted little attention. After all, the many stars in the Milky Way galaxy do not seem to be moving away from each other, but instead to be traveling in circular orbits around its massive central region. But in 1929, Edwin Hubble, then an astronomer at the Mount Wilson Observatory, announced that the galaxies he had been observing were in fact receding from us, and from each other, at speeds of up to several thousand kilometers per second. These galaxies, like the Milky Way, apparently keep their internal structure intact over long periods of time; individually they are moving through space, more or less as units, or "particles." The Einstein theory could be applied to galaxies instead of to stars.

RECEDING GALAXIES

To determine the speeds of galaxies, Hubble made use of the Doppler effect. The Doppler effect is the phenomenon experienced when a source of waves, such as light or sound, is moving with respect to an observer or listener. If the source approaches a person, the latter perceives the waves as rising in frequency: sound becomes higher-pitched or light tends toward the violet end of the spectrum. If the source is moving away from the person, the waves drop in frequency: sound becomes lower in pitch or light tends toward the red end of the spectrum. In examining the light

from the galaxies spectroscopically, Hubble noted that the lines were shifted out of their customary positions, toward the red end of the spectrum. He concluded that this was due to the motion of the galaxies away from the earth. Moreover, the greater the speed of recession, the greater was the shift toward the red—the so-called "red shift."

Hubble was able to deduce the velocities of the receding galaxies from the amount of red shift. He determined that a shift of one per cent toward the red corresponded to a speed of 3,000 kilometers per second, for example. Furthermore, the apparently fainter and therefore more distant galaxies seemed to have greater red shifts—i.e., they were traveling the fastest of all. More recent work on the problem has disclosed that there is an increase of speed of about 32 kilometers per second for each million light-years of distance outward into space.

A light-year is a unit of measurement used in astronomy. It is the distance that light travels in a year. Since light travels at 300,000 kilometers a second, it goes about 9,500,000,000,000 kilometers in a year. Thus, when a galaxy is 3,000,000 light-years from the earth, it is traveling about 95,000 kilometers a second. However, recent studies, which indicate that some objects in the universe may be traveling at higher than previously calculated speeds, have thrown into question the significance of red shift as an indication of speed.

Such figures represent the distances and speeds of galaxies as they appear to us now. We see the galaxies as they looked hundreds or thousands of millions of years ago, since their light has taken that long to reach the earth. Although the figure of 32 kilometers a second for each million light years seems to be constant, there is some reason to believe that it is changing very slowly with the passage of time.

The big-bang theory is also based on the observation that galaxies are distributed more or less uniformly throughout space, much like a cloud of particles. The physicist George Gamow has explained the expansion effect by an analogy with an inflating balloon. If one blows up a balloon that is

uniformly covered with specks of paint, one sees all the specks move away from each other.

As the specks or galaxies move away from each other, observers on any one of the specks (as persons on earth in the Milky Way) would get the impression that all the other specks (galaxies) were moving away from them. This impression of outward expansion for all observers, in fact, has been termed the *cosmological principle*.

STEADY-STATE THEORY

Back in 1948, however, there was not sufficient information available to test the big-bang theory. British astronomer Fred Hoyle and some British astrophysicists argued for another theory, the steady-state theory, according to which the universe not only is uniform in space—the cosmological principle—but also unchanging in time—the perfect cosmological principle. Thus, the cosmological principle was extended so that it was “perfect” or “complete” and not conditioned by specific historic events. The steady-state theory is in direct conflict with the big-bang theory. In the latter theory, space becomes progressively emptier as the galaxies recede from one another. In the steady-state theory, one must postulate that new matter is continuously created in the space between the galaxies, so that new galaxies may then form to take the place of those which have receded. The new material is believed to be hydrogen, which is the source from which stars and galaxies spring.

Continuous creation of matter from apparently empty space has met with considerable skepticism from authorities, because it seems to violate one of the basic laws of physics—the conservation of matter. Matter cannot be created or destroyed,

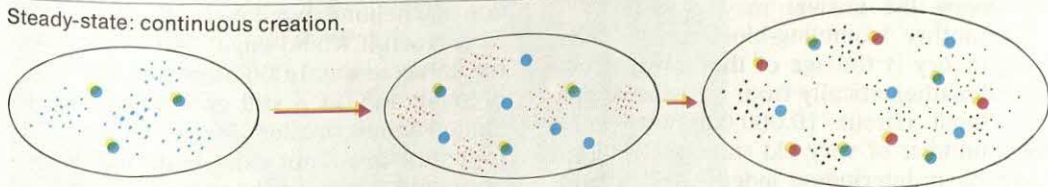
but just converted into other kinds of matter or into energy. On the other hand, it would be difficult to disprove continuous creation directly, because the amount of matter, according to steady-state theory, is increasing very slowly—about one atom per thousands of millions of years in a volume of space equal to that contained in an average television set.

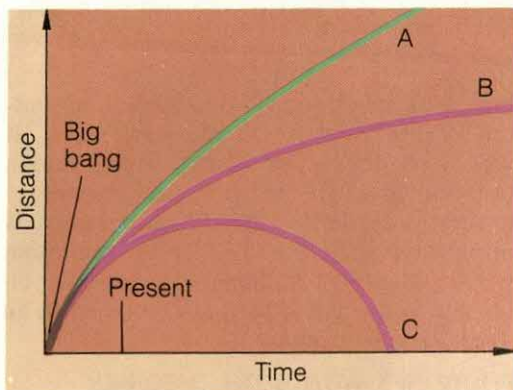
PROBLEMS WITH BOTH THEORIES

The steady-state theory in its original form has not fared well in the light of recent astronomical studies. Its chief failing appears to be its insistence on the uniformity of stars and galaxies. It did not predict that the average properties of nearby and distant galaxies would be different. And yet differences have been observed, especially by radio astronomers. There are far more faint radio sources than have been predicted by the steady-state theorists. Such effects, however, can be explained on the basis of big-bang theory, since it predicts that galaxies evolve with the passage of time. Because of the finite speed of light, we see various galaxies as they appeared at different stages in the past, as we have already pointed out. For very distant galaxies, thousands of millions of light-years off, we see them as they were thousands of millions of years ago, when their characteristics were different. Because of these facts, Hoyle and others have renounced the early form of the steady-state theory. Others are attempting to modify it, however.

The big-bang model has also run into difficulties. For one, the concept of all matter being compressed into one dense mass has been very difficult for present-day physics to describe or comprehend. The explosion and consequent motion of this material may be compared to the possible

Steady-state: continuous creation.





Three possible ways in which the big-bang universe could be evolving. Curve A represents a rapidly and continuously expanding universe. Curve B shows a universe expanding slowly. Curve C shows a universe where expansion has stopped, overcome by the force of gravitation, and contraction is occurring.

flights of a projectile fired off the earth's surface. It may travel so fast that it easily escapes the earth's gravity and moves very quickly far out into space, and keeps doing so. Or it may just barely escape the earth and keep moving very slowly on outward. Or it may never have had sufficient velocity to rise free of earth's gravity. The projectile will eventually slow down, stop, and then start falling back to earth, at increasing speed. The expanding universe may be acting similarly—expanding rapidly, or comparatively very slowly or finally so slowly that expansion may halt, being overcome by the gravitational attraction of matter for matter. Then contraction sets in, as all matter begins to “fall” back, faster and faster, toward its point of origin. Some astronomers think that ultimately, the whole process may begin again, after matter has become sufficiently compacted to explode once more. At the present time, observations have not been good enough to suggest which one of the three expansion processes, described above, may be operating in the universe. Yet, there is hope that in a few years the answer may come.

Another stumbling-block to the big-bang theory is the age of the universe obtained mathematically from it. It seems just too small (about 10,000,000,000 years) for a number of very old stars, whose ages have been determined independently from the amount of nuclear fuel they are sup-

posed to have consumed. However, proponents of the big-bang theory feel that this method of stellar dating is not yet very accurate, and so they consider that it is not really a serious threat to their idea.

ONCE A SMALL UNIVERSE?

The present known proportion of helium in the sun and in many other stars averages about 20 to 30 per cent by weight. This agrees with the amount predicted by the big-bang theory. Big-bang exponents say that the temperature of the cosmic blast first reached several thousand million degrees, but began to drop off after the event. At some stage many atomic nuclei began to form, including those of helium. The exact amount of early helium depends on several factors, particularly the temperatures prevailing then and now. Some trouble has been experienced in the theory in this regard. Critics also point out that the apparent agreement of present helium abundance between fact and theory may be only coincidence, and that later cosmic events influenced the helium concentration.

The explosive fireball must have given off terrific amounts of high-energy radiation. Gamow and others said that remnants of this radiation might still be detectable, but now at much lower frequencies, because of the great loss of heat and energy since the big bang. In 1965 and later, such possible radiation was detected, in the microwave region of the spectrum, at various wavelengths. It seems to come evenly from all regions in space, as would be expected from a universal fireball, and corresponds to a present temperature of from 2.5 to 3° above absolute zero, or about -270° Celsius. These observations are in very close agreement with the predictions that have been made by theory.

If all these findings about fireball radiation and helium abundances finally prove to be correct, it would mean that at one stage the universe was 1,000 times smaller than it is today and, at a still earlier period, 300 million times smaller. Matter in this compact state could not exist as normal atoms, nor could it eventually come to form stars and galaxies, unless it exploded.

THE NIGHT SKY

by Marian Lockwood

The pageant of the night skies aroused the curiosity of men from the very earliest times. As observations piled up and astronomical instruments were perfected, the pattern of the sky became clearer. Though there is still a great deal that we do not know about our universe, we now have a tolerably accurate idea of the face of the sky as it is viewed from our planet, the earth.

Of the countless thousands of millions of celestial bodies, we can see either with the naked eye or the optical telescope only those bodies that shine. Some of the heavenly bodies—the stars, for instance, which are suns like our sun—shine by their own light. Others, like the planets and their moons, shine because they reflect the light of the sun, as a mirror reflects the light that shines upon it. The earth, because it reflects the sun's light, would also appear as a shining body to a nearby observer.

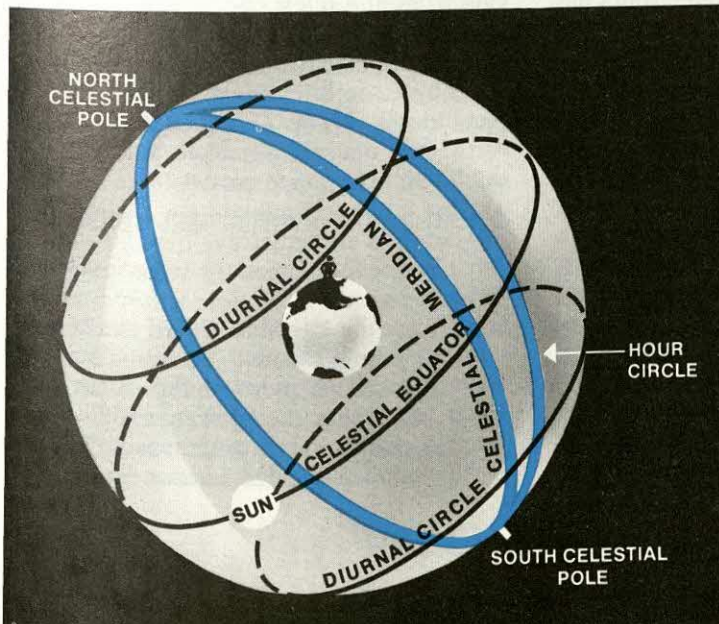
The stars make up by far the largest part of the individual objects that we observe in the sky. The average person looking at the sky can see perhaps 4,000 stars at one time. Other naked-eye stars, numbering perhaps 5,000 more, would not be visible to him. For one thing, we see only one-half of the sky at any one time. Besides,

some of the stars that are close to the horizon are lost in the haze that so often obscures the sky just above the horizon line. The astronomer with a powerful array of telescopes can see or photograph countless individual stars and star groups that are not visible to the naked eye.

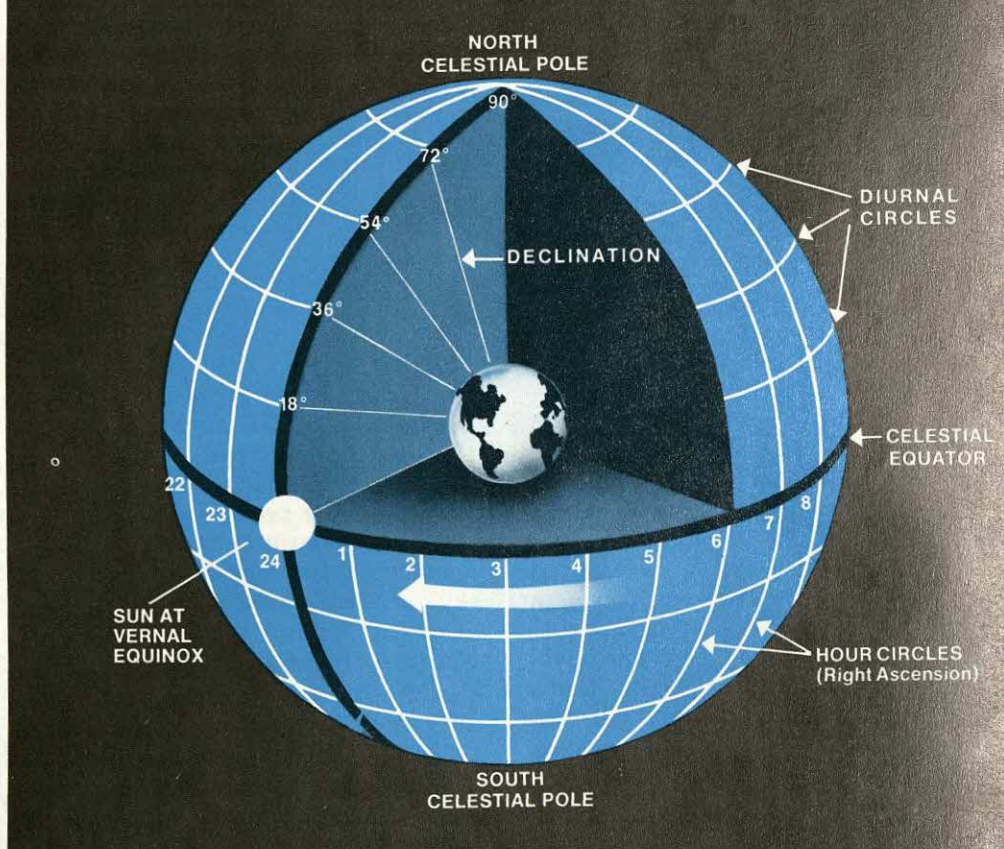
Long ago, possibly before the dawn of history, people began to imagine that they could see patterns or designs in the grouping of stars. One saw a bear outlined in the sky, another a winged horse. Still others imagined that they beheld the likenesses of various mythological heroes. And so, through the ages, the stars became divided into groups, or constellations, of which there are now eighty-eight. The whole sky is filled with these star groups. There are no empty spaces left between the constellations, and they do not overlap. Each star belongs in a specific constellation and each constellation has its own name.

THE SUN AND PLANETS

The star with which we are most familiar is the sun—the center of our solar system. It provides the light and heat that are essential for life upon the earth. Our sun is only an average star in size and brightness, but from our point of view it is the brightest



The celestial sphere with the lines of reference used in the equatorial system indicated. Diurnal circles mark a star's declination, or position north or south of the celestial equator. Hour circles mark a star's east-west position, or its right ascension.



A star's east-west position, or right ascension, is measured eastward from the vernal equinox to the place where the star's hour circle crosses the celestial equator. The measurement is given in degrees, minutes, and seconds, with one hour being equal to 15 degrees. The vernal equinox is one of the two places where the celestial equator crosses the ecliptic, or apparent path of the sun.

and most conspicuous object in the sky. Like the other stars, the sun is a large mass of self-luminous gas, but to us it looks like a solid sphere with a sharply defined disk, across whose surface spots travel from time to time. Above the disk there is an atmosphere made up of several layers.

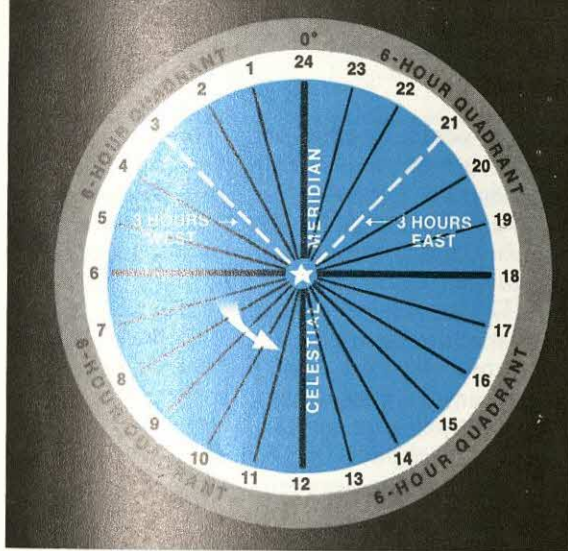
Next to the sun, the moon is the brightest object in the sky, although actually, of course, it shines only by the reflected light of the sun. The moon is the satellite of the earth, revolving around the earth in an elliptical (oval) orbit. The moon has a constantly changing appearance as it passes through its various phases.

The earth travels through the universe in company with eight other planets. They all move in elliptical orbits around the sun. The planet Mercury is nearest the sun; then comes Venus; then the Earth. Beyond the Earth are Mars, Jupiter, Saturn, Uranus,

Neptune, and Pluto. All the planets except Uranus, Neptune, and Pluto are at certain times visible to the naked eye. Even Uranus can be dimly seen at infrequent intervals without a telescope, if one knows just where to look for it. The earth is not the only planet that has a satellite. Jupiter has 16 known satellites, Saturn 17, Uranus 5, Neptune 2, Mars 2, and Pluto, like the earth, has one. This makes a total of 44 satellites for the nine planets.

OTHER SOLAR SYSTEM MEMBERS

Certain features of the solar system are to be seen only from time to time, greatly enlivening the celestial show as they become visible. Meteors—"shooting stars"—sometimes flash across the night sky. Each represents the glow caused by friction as matter from outer space hits the earth's atmosphere. At various times of the



Hour circles emanating from the celestial north pole. All hour circles run through the celestial pole and cross the celestial equator at right angles. There is no limit to the number of possible hour circles that can be drawn, and each object in the sky has its place on one of these lines.

year, the earth, in its journey around the sun, passes through large crowds of meteoritic particles. Then hundreds and even thousands of "shooting stars" stab the blackness of the night with their quick, sudden trails of light. There are a number of meteor showers in the course of a year.

Comets, much more infrequent visitors than meteors, proceed majestically across the sky and may be visible for many nights or even for months. They move in greatly elongated paths around the sun. Some comets come back to the sun's neighborhood every few years, while others complete their journey only in many centuries. Comets often develop luminous tails, hundreds of thousands of kilometers long, when they approach the sun.

To complete our rapid sketch of the solar system we should mention here the magnificent displays called the auroras: the aurora borealis, or northern lights, in the Northern Hemisphere; the aurora australis, or southern lights, in the Southern Hemisphere. Caused by electromagnetic disturbances on the sun that toss electrons into the earth's magnetic field, auroras glow high in the atmosphere and form shafts of shimmering colored lights.

BEYOND THE SOLAR SYSTEM

Far out and beyond our miniature universe of sun and planets lies the Milky Way, a majestically beautiful pathway of light girdling the heavens. Although it appears to be cloudy, it is actually composed of suns like our own. It forms part of the Milky Way galaxy, our island universe. The Milky Way is the galaxy to which our sun belongs. It is one of countless galaxies in the universe. The Milky Way is shaped like a wheel with a central hub. When we look in the direction of the Milky Way, we are looking, from our position near the sun, into the center of the galaxy. That is why the stars appear so close together and suggest a luminous band. The individual stars in our night sky also belong to the Milky Way system. They seem scattered because when we see them we are looking towards the outskirts of the galaxy.

THE CELESTIAL SPHERE

Such, then, are some of the objects that we see in the skies. To locate these objects in space, astronomers have developed the concept of a celestial sphere. The celestial sphere is a sphere extending to the outermost limits of space with the earth at its center. On this imaginary sphere, they have drawn reference lines. By referring to them, they locate heavenly bodies, plot their paths, and keep track of eclipses.

Geographers have done much the same thing in drawing their imaginary reference lines on the surface of the earth. We locate a city on the map by latitude and longitude. The latitude is the city's position north or south of the equator—an imaginary line, drawn around the earth equally distant from the North and South poles. The city's longitude is its position east or west of another imaginary line, known as the zero, or Greenwich, meridian. This line, which is the starting point for all longitude, passes through the North and South poles of the earth and through Greenwich, England. This line is also the starting point for the standard time zones.

All the imaginary lines of reference in the sky, as we shall see, are either circles or

arcs—parts of circles. The circumference of any circle, large or small, can be divided into 360 degrees (360°). Each degree is divided into sixty minutes ($60'$) and each minute into sixty seconds ($60''$). Because the distance between any two celestial bodies is measured along the arc of a circle, we measure sky distances in terms of degrees, or minutes, or seconds of arc. For example, a star halfway up the sky from the eastern horizon has an altitude (height) of 45 degrees of arc.

The equatorial system. Several systems are employed in establishing reference lines on the imaginary sphere. The equatorial system is the one generally used by astronomers. The system is defined with reference to the earth's axis. Imagine this axis as a long rod extending out into space until it touches our imaginary celestial sphere. The point where the north end of the rod touches the sphere is the north celestial pole. The bright star Polaris ("Pole Star") is close to this point in the heavens; hence it is often called the North Star. Its technical name is Alpha Ursae Minoris, which means in the astronomer's language "brightest star in the constellation Ursa Minor." The south end of the rod touches the celestial sphere at the south celestial pole. There is no star in its immediate vicinity. Equidistant between the celestial poles and encircling the heavens is the imaginary line known as the celestial equator. It is really an extension of the earth's equator into space. The celestial equator divides the sphere of the heavens into halves—the Northern and Southern Hemispheres.

Running through the celestial poles and at right angles to the celestial equator are great circles known as *hour circles*. There is no limit to the possible number of hour circles that could be drawn. Each object in the sky has its place on one of these circles.

A star's *declination* is its position—given in degrees, minutes, and seconds—north or south of the celestial equator. Another way of describing declination is to say that it is a star's angular distance north or south of the celestial equator on its particu-

lar hour circle. North declination is marked plus; south declination, minus. A star that is 10 degrees north of the celestial equator is said to have a declination of $+10^\circ$. One that is 49 degrees south of the equator has a declination of -49° .

The east-west position of a star is known as its *right ascension*. It is always measured eastward from the point known as the *vernal equinox*, one of the two places where the celestial equator crosses another circle, the ecliptic. The ecliptic marks the path of the sun's apparent annual journey through the heavens. It is inclined about $23\frac{1}{2}$ degrees to the celestial equator. The right ascension of a star is measured from the vernal equinox to the place where the star's hour circle crosses the celestial equator. It is usually given in terms of hours, minutes, and seconds rather than in degrees, minutes, and seconds. One hour equals 15 degrees; 24 hours equal 360 degrees.

The horizon system. In the horizon system, all positions are referred to the horizon—the line where the earth seems to meet the sky. When an astronomer speaks of the horizon, he does not have in mind the irregular, broken skyline that we usually see. The astronomical horizon is known as the *true horizon*. It is what you would see from the middle of the ocean—a circle stretching unbroken all around you and at the same level everywhere.

The point on the celestial sphere directly above you is the *zenith*. The great circle passing through the zenith of a given place and also through the two celestial poles is called the *meridian* of that place. Directly beneath you, on the celestial sphere on the other side of the earth, is the *nadir*. As you change your position on the earth's surface, your zenith and your nadir move with you. No one else can have the same zenith and nadir as you at a given time.

In the horizon system, the position of an object in the heavens is described by its *altitude* and its *azimuth*. The altitude is measured in degrees, minutes, and seconds from the horizon up toward the zenith. Points below the horizon are indicated by negative altitudes. The azimuth of a body in

Date 26-12-88

Rec. No. 4813

space is measured in degrees, minutes, and seconds westward, clockwise, along the horizon from the south point of the horizon to the foot of the vertical circle of the body. The vertical circle is an imaginary circle passing through both the body and the zenith. Navigators and surveyors measure azimuth from the north point of the horizon through the east.

In the horizon system, the position of a heavenly body will vary according to the place from where it is observed. Hence position and time must be given.

Other systems. The ecliptic system is based on the ecliptic. It is used to describe the position of members of the solar system, most of which move in or near the ecliptic. The galactic system, used to study the motions of stars, is based on the plane of our galaxy. We compared the latter previously to a wheel. We can think of the plane of the galaxy as cutting the wheel into upper and lower halves.

THE CHANGING FACE OF THE SKY

The earth spins completely around on its north-south axis about once in 24 hours. As it does so, its eastern horizon dips down, and we see new stars coming into view in the eastern part of the sky. At the same time, the western horizon rises, and the stars that we have been observing low in the west are gradually hidden from our sight. This accounts for the apparent rising of stars in the east and their setting in the west. The sun, the moon, and the planets—and comets, when they happen to be visible—also make their first daily appearance from below the eastern horizon.

But all these heavenly bodies are not actually moving around the earth from east to west. This is an optical illusion, caused by the spin of the earth from west to east. However, there are exceptions to this illusion. This depends on the latitude from which you are observing the night sky. And it also depends on whether the stars are circumpolar, which is explained later.

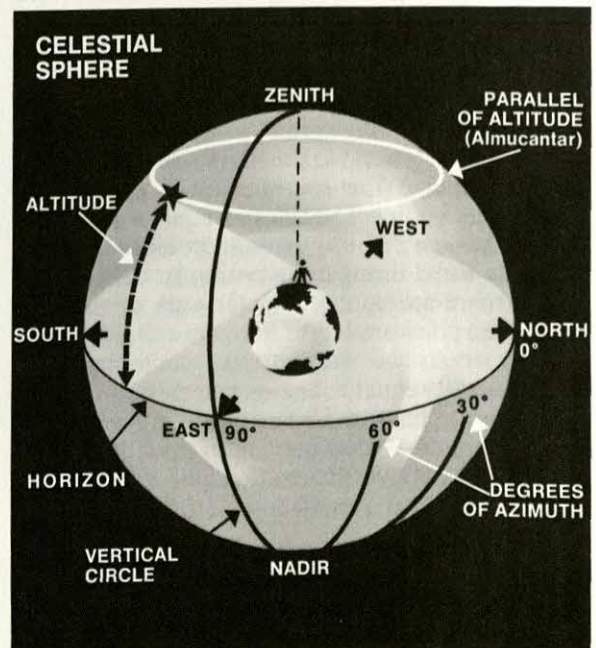
The face of the sky changes not only with the turning of the earth but also as we travel north or south on the earth's surface.

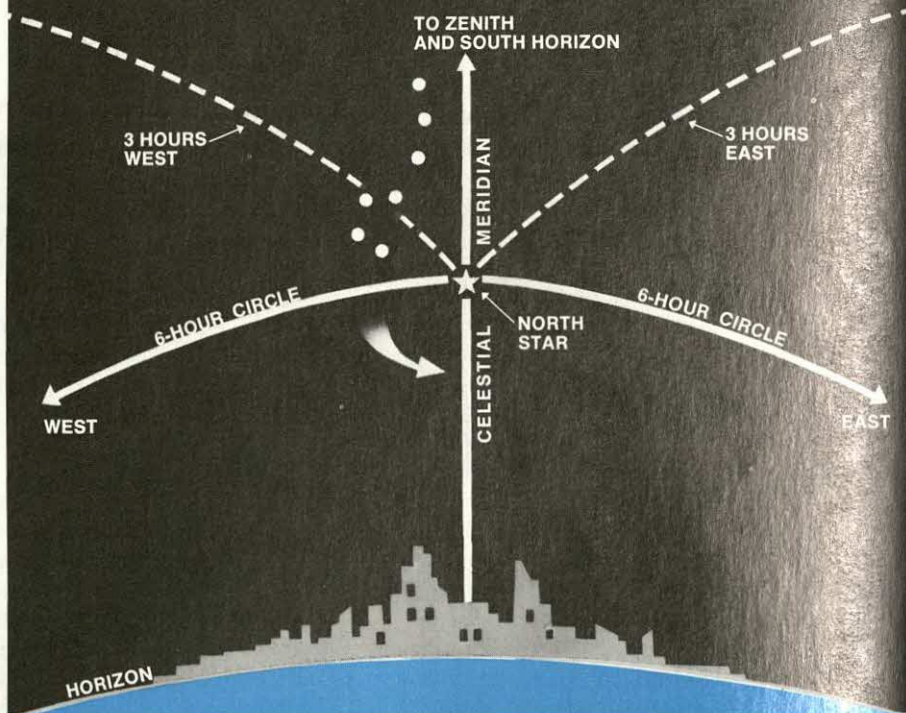
A person would never see a rising or setting star at the North Pole. There the stars seem to move around the sky in circles parallel to the horizon. This aspect of the heavens is called the *parallel sphere*. At the North Pole, you would be able to see the entire northern half of the celestial sphere—the Northern Hemisphere. If you were at the South Pole, you would see the entire Southern Hemisphere of the sky. At either pole, the celestial pole is directly overhead and coincides with the zenith.

As we start traveling south from the North Pole, we leave behind us the parallel sphere, where all the stars revolve in circles parallel to the horizon. If we stop 40 degrees south of the pole, at latitude 50, we find that the north celestial pole is now 50 degrees above the northern horizon. As we watch the northern night sky from this vantage point, we see that a good many stars, circling the pole, never seem to set. These are known as the *circumpolar stars*.

If, from latitude 50 degrees, we look to the south, we see stars rising and setting. They come over the eastern horizon at an oblique angle and they set obliquely in the

The celestial sphere with the lines of reference used in the horizon system indicated. In this system, all points in the sky are located in terms of the observer's horizon. Therefore the observer's position and time must be given.





One way to map the sky is shown above and on the opposite page. First locate the celestial meridian, or great circle passing through the zenith and through the celestial poles. Then face the North Star and point to it with your right hand. Then swing your arm sideways and to the right until it is level with your shoulder and due east. Do the same thing with your left arm. The imaginary lines you thus locate are the 6-hour circles; they divide the sky around the North Star into quadrants. Then trace a path in the sky at right angles to the 6-hour circles. This line is the celestial equator. You can check your determination of the celestial equator if the constellation Orion is visible, for the right hand star in Orion's belt is practically on the celestial equator.

west. The aspect of the sky as seen from any point between the poles and the equator is known as the *oblique sphere*.

When we observe the heavens from the equator of the earth, we see the so-called *right sphere*, another aspect of that ever-changing sky above us. Here, the stars rise vertically and set vertically—at right angles to the horizon. The north celestial pole is located at the north point of the horizon, and the south celestial pole at the south point. The celestial equator runs from the east point to the west point of the horizon and through the zenith. At the equator, there are no circumpolar stars at all, since the poles are on the horizon.

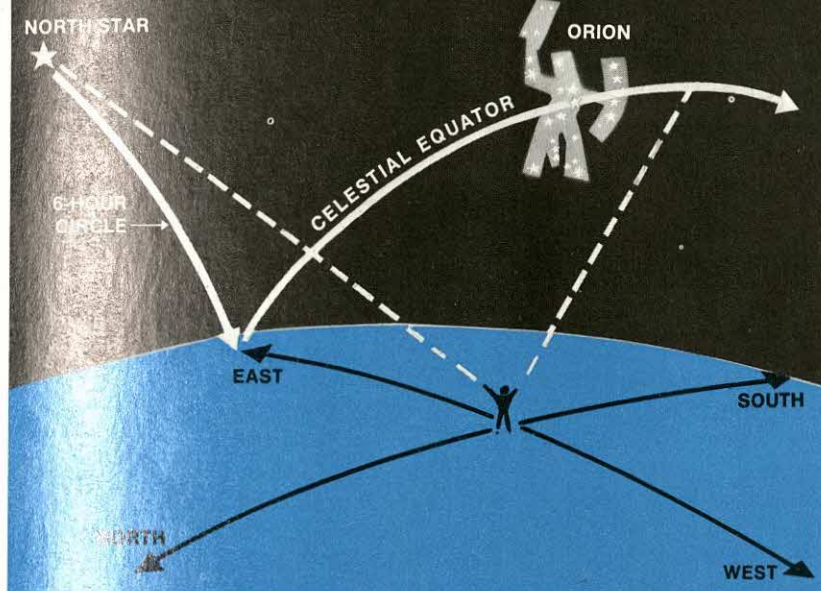
As we travel to the south from the earth's equator, the south celestial pole rises gradually over our southern horizon. At latitude 10 degrees south, it will be 10 degrees above the southern horizon; at altitude 80 degrees south, 80 degrees above

the southern horizon and so on. The area of circumpolar stars increases as we go south. At the South Pole of the earth, we again observe the parallel sphere.

THE SUN'S APPARENT JOURNEY

Like the other celestial bodies in the heavens, the sun seems to be constantly moving. Not only does it apparently rise and set each day, but it seems to be changing its position with respect to other stars.

To understand why this is so, let us consider again the earth on which we live. In addition to turning on its axis about once every twenty-four hours, it also moves around the sun at a speed of about 29 kilometers a second. Its path around the sun is elliptical, and it takes our planet about a year to complete its journey. As we travel around the sun, we see that bright body first against one part of the sky and then against another. And since we are not conscious,



from any evidence of our senses, of the fact that the earth is moving around the sun, the sun seems rather to us to be moving around the sky to the east.

How can we detect this apparent eastward journey of the sun through the heavens? It is true that in broad daylight, we never see the stars, since the sun's light blots them out. However, if we watch in the early evening just after the sun has set and in the dawn before sunrise, we can see at some time or other a great many of the stars that are hidden by the sun during the day. In the evening—sometimes even in the twilight—we can briefly glimpse the brighter stars that are just to the east of the sun and that set soon after it goes down. In the morning, in the same way, we can see stars just to the west of the sun before the sun comes up. The stars that become visible in this way are always changing because of the sun's steady eastward motion.

In the course of one year, the sun completes its circuit. This apparent path of the sun is a great circle of the celestial sphere—360 degrees. Since our year is about 365 days long, the sun appears to move toward the east approximately one degree each day.

THE EQUINOXES

As we have pointed out, the apparent yearly path of the sun around the heavens is known as the *ecliptic*. It is one of the most important reference points in the sky. In-

clined at an angle of $23\frac{1}{2}$ degrees to the celestial equator, the ecliptic crosses it at two points 180 degrees apart. These two points where ecliptic and equator meet are known as the *equinoxes*. The word "equinox" means "equal night." When the sun is at either of these points, our day is equal to our night in length. The point where the sun crosses the celestial equator about March 21 is known as the vernal equinox. The point where it crosses the equator about September 23 is the autumnal equinox.

Since the ecliptic crosses the equator at two points, the sun in its apparent journey around the earth spends half of the year in the northern celestial hemisphere and half in the southern. Each day its right ascension and its declination change by a very little. If you have ever, over a period of time, watched the sun rise and set, you have noticed that it does not do so at exactly the same place on the horizon day after day. At the equinoxes, the sun rises and sets exactly at the east and west points of the horizon. At that time, the days and nights over most of the earth are equal in length. From the time of the vernal equinox until about June 21, the sun follows the ascending line of the ecliptic in the Northern Hemisphere. During this time, it rises and sets farther north each day. For those who live in the Northern Hemisphere, this makes the days increasingly longer. About June 21—the longest day of the year in the Northern Hemisphere—the sun reaches its

most northerly point in the noon sky— $23\frac{1}{2}$ degrees north of the celestial equator—and its most northerly point on the horizon. This point in the sky is called the *summer solstice* and is just above the earth's Tropic of Cancer, which is $23\frac{1}{2}$ degrees north of the earth's equator.

Then the sun appears to slide down the ecliptic toward the celestial equator and the days begin to shorten. About September 23, the sun's path again crosses the equator and then continues on its southern course. For those living in the Northern Hemisphere, the nights grow longer than the days. About December 21, the sun reaches its most southerly point—the *winter solstice*—and the shortest day of the year for inhabitants of the Northern Hemisphere has come. This southernmost point is above the earth's Tropic of Capricorn, $23\frac{1}{2}$ degrees south of the equator.

The ecliptic is the central line of a band in the heavens known as the *zodiac*—a band 16 degrees wide (8 degrees each side of the ecliptic). It is divided into twelve parts, or signs, of 30 degrees each. They are named for the twelve constellations through which the sun moves. The zodiac has no significance in astronomy.

TELLING TIME

The sun is an important factor in the telling of time because of its apparent year-long journey around the sky and diurnal (daily) motion. The sundial was the earliest timekeeping instrument.

The time told by the sundial is *apparent solar time*, based on the movement of the sun as we see it. But the sun sometimes seems to proceed faster (or slower) in its slight daily movement toward the east than it does at other times. This makes it a bit unreliable as a clock. To make up for this natural deficiency, astronomers have invented what they call the mean sun, or average sun. Its speed is an average figure obtained by taking the average of the real sun's speeds during the year. The mean sun is the basis of *mean solar time*, which is used for ordinary purposes.

Astronomers use another kind of time—*sidereal time*, or star time. It is based

on the sidereal day—the interval between two crossings of a given star across the meridian. It is shorter than the day marked out by the sun. Any star (except the sun) comes back to the meridian in 23 hours, 56 minutes, and 4.09 seconds. This is the actual period of time in which the earth spins once. Our ordinary solar day, however, is 24 hours long—almost four minutes longer than the sidereal day. The lag is due to the sun's slipping back toward the east each day in the course of its apparent journey through the sky. This means, of course, that sidereal hours, minutes, and seconds are shorter than the hours, minutes, and seconds of the solar day.

OTHER CHANGES

The stars appear to be firmly fixed in their places. Actually, however, they have their own particular type of motion in space. This is called *proper motion*—motion across the line of sight, neither toward us nor away from us. As a result of their proper motion, the stars gradually change their positions with respect to one another. This comes about so slowly, however, that hundreds or thousands of years may elapse before we can detect the altered pattern.

Because of gradual changes in the relative positions of stars, the constellations just as gradually change form. For example, the constellation called Ursa Major (Great Bear), or the Big Dipper, actually looks like a dipper today. Some 50,000 years from now, it will resemble a flat frying pan.

Another striking change in the face of the sky is caused by what is called *precession*. This represents the wobbling of the earth on its axis. As a result of the earth's wobbling, the North Pole traces out a great circle in the skies—a circle that is completed in a little less than 26,000 years. As the North Pole makes this circuit, the North Star is constantly changing. Right now, as we have observed, the star Polaris is closest to the North Celestial Pole and is our North Star. In about 12,000 years, Vega, in the constellation Lyra, will be the star nearest the pole. After the completion of the precession cycle, Polaris will again be the North Star.



American Museum of Natural History

A mid-19th century U. S. artist's conception of some of the winter constellations as seen from middle northern latitudes.

THE CONSTELLATIONS

by Marian Lockwood

A walk in the woods is more interesting if we can recognize the birds and their songs and name at least some of the wild flowers and trees. A knowledge of the stars can, in the same way, add to our enjoyment of the night sky. But we can succeed in knowing them only if we take the time, night after night and season after season, to observe them closely.

As you must have noticed, some stars appear to be brighter than others. We cannot, however, tell how bright a star really is just by looking at it. A brilliant one that is far away may seem much fainter than a less brilliant one that is nearer. The astronomer calls the brightness of a star as it appears to us its apparent magnitude or, simply, its magnitude. A first-magnitude star is about

2½ times as bright as a second-magnitude star, which in turn is about 2½ times as bright as one of third magnitude and so on. The few stars that are brighter than standard first magnitude are given minus magnitudes. A list of symbols for different star magnitudes is given with Map 1.

The sky is divided into eighty-eight (some astronomers list eighty-nine) star groups, or constellations. Each constellation has its own name and definite boundaries. The most conspicuous ones were named by stargazers several thousand years ago. Some have received names in comparatively modern times.

Some individual stars—usually the brighter ones—have also been given names, such as Sirius and Aldebaran. Many are

designated by letters of the Greek alphabet followed by the Latin name of the constellation. For example, Sirius, the brightest star in the constellation Canis Major (the Greater Dog) and the brightest star in the whole sky, is also known as Alpha (α) Canis Majoris. In most cases, though not in all, the brightest star in a constellation is the alpha of that group, the next brightest is beta, the next gamma and so on, down through the Greek alphabet. The letters of the Greek alphabet are given with Map 1.

READING A SKY MAP

To locate the constellations, you must know how the stars and the sky change position night by night and season by season as well as with change of latitude. The four full-page star maps reproduced in this book show the heavens as they appear from middle northern latitudes. In using the maps, remember that any given star comes back to the same place in the heavens about four minutes earlier each night. This makes a difference of about two hours in one month. A star that rises at 10 P.M. tonight will rise about 8 P.M. a month from now.

Map 1 shows the main northern constellations that are visible all night long every night of the year from Lat. 40° N. (40° North Latitude) northward. They never disappear below the horizon.

To use Map 1, find the approximate date in the outer circle. If you are observing on November 15, for instance, and facing north, hold the map so that the middle of the section labeled November is at the top. Your meridian—the imaginary line running from the north point of your horizon through the zenith, or the point directly overhead, to the south point of your horizon—will then run exactly from the top of the map to the bottom. The stars on that line will be on your meridian about 9 P.M. standard time, whatever your longitude. Those at the bottom of the map will be close to the northern horizon and those at the top will be higher in the sky than Polaris, the star that almost exactly marks the north pole—the central point of the northern heavens. The figures in the circle within the outer rim mark the hours of right ascen-

sion, which tell you how far east or west a star is in the sky.

You may find it helpful, in orienting yourself in relation to the sky, to hold the map over your head, as you face the north. Since the stars at the bottom of the map are those close to the northern horizon, hold the bottom toward the north and the top toward the south.

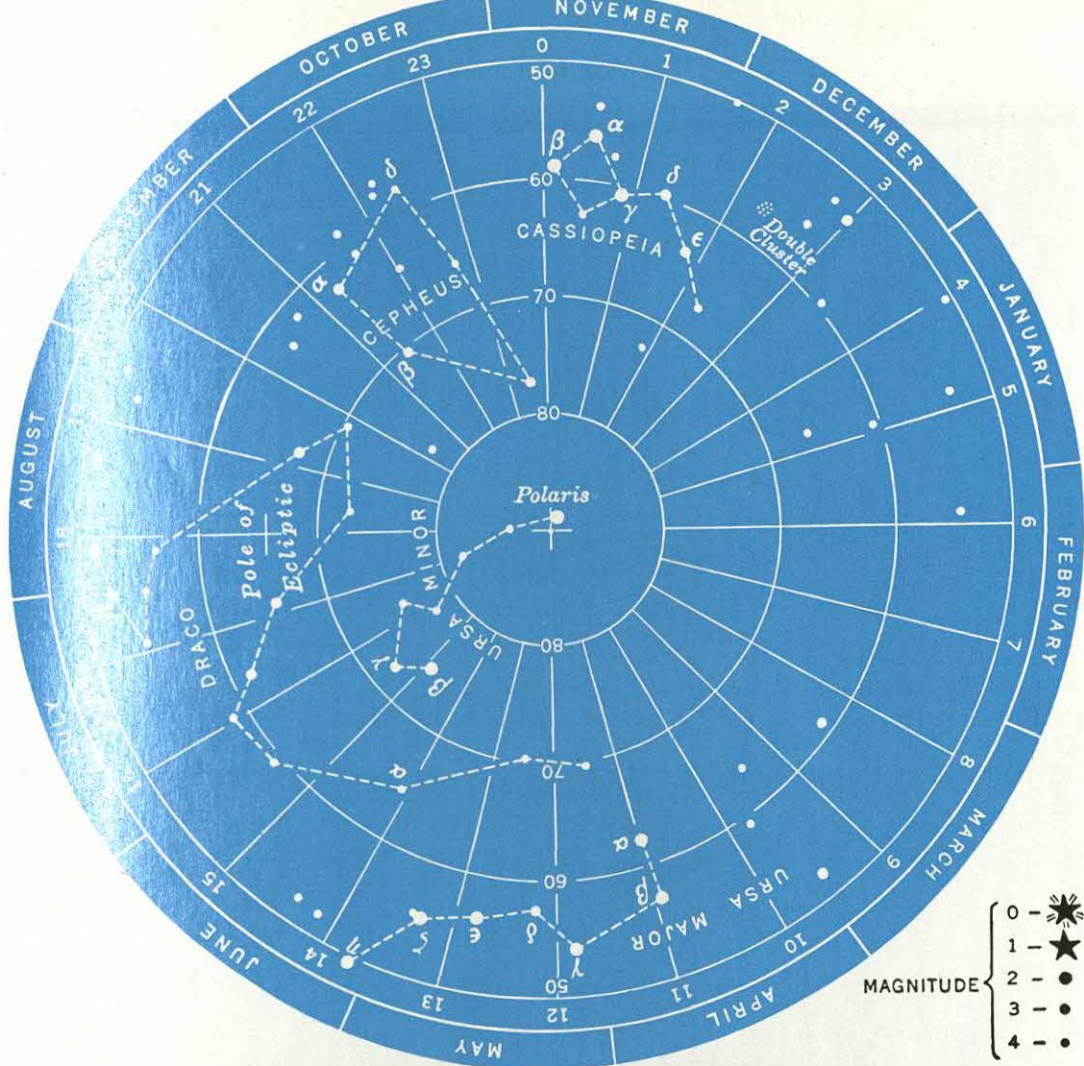
Suppose you want to find the stars that will be on your meridian later—say 11 P.M.—on that same night, November 15. Turn the map counterclockwise from the position for 9 P.M., through two hours of R.A. (right ascension). The top of the map will then be about halfway between R.A. 2 and 3. Again, the stars at the bottom of the map will be near the northern horizon and those at the top, above the pole. If you want to observe at 7 P.M. on the same night, turn the map through two hours of R.A. in a clockwise direction.

The concentric circles in the map show you the declination of the stars. The declination of a star is its angular distance from the celestial equator. It will help you to estimate sky distances in degrees if you remember that from the north point of the horizon to the south point is 180°, or half a circle. From the horizon to the zenith is 90°.

NORTHERN CONSTELLATIONS

Perhaps the most familiar group of stars in the entire sky is the Big Dipper, which is part of the constellation Ursa Major, or the Greater Bear. You can easily find the Big Dipper at the bottom of Map 1, clearly outlined by its seven bright stars. If you draw a line through Beta and Alpha (the pointers) at the outside of the Dipper's bowl, and extend that line about five times, you will find Polaris.

Polaris is approximately over the earth's north pole and is the central point around which the stars of the Northern Hemisphere turn in a counterclockwise direction. Polaris will always show you where the north is. The star is at the end of the handle of the Little Dipper, which you can see "hanging" from Polaris. The Little Dipper, in its turn, belongs to Ursa Minor, the Lesser Bear.



From *Astronomy* by Robert H. Baker, 8th ed., D. Van Nostrand Co., Inc.

Map 1. The northern constellations

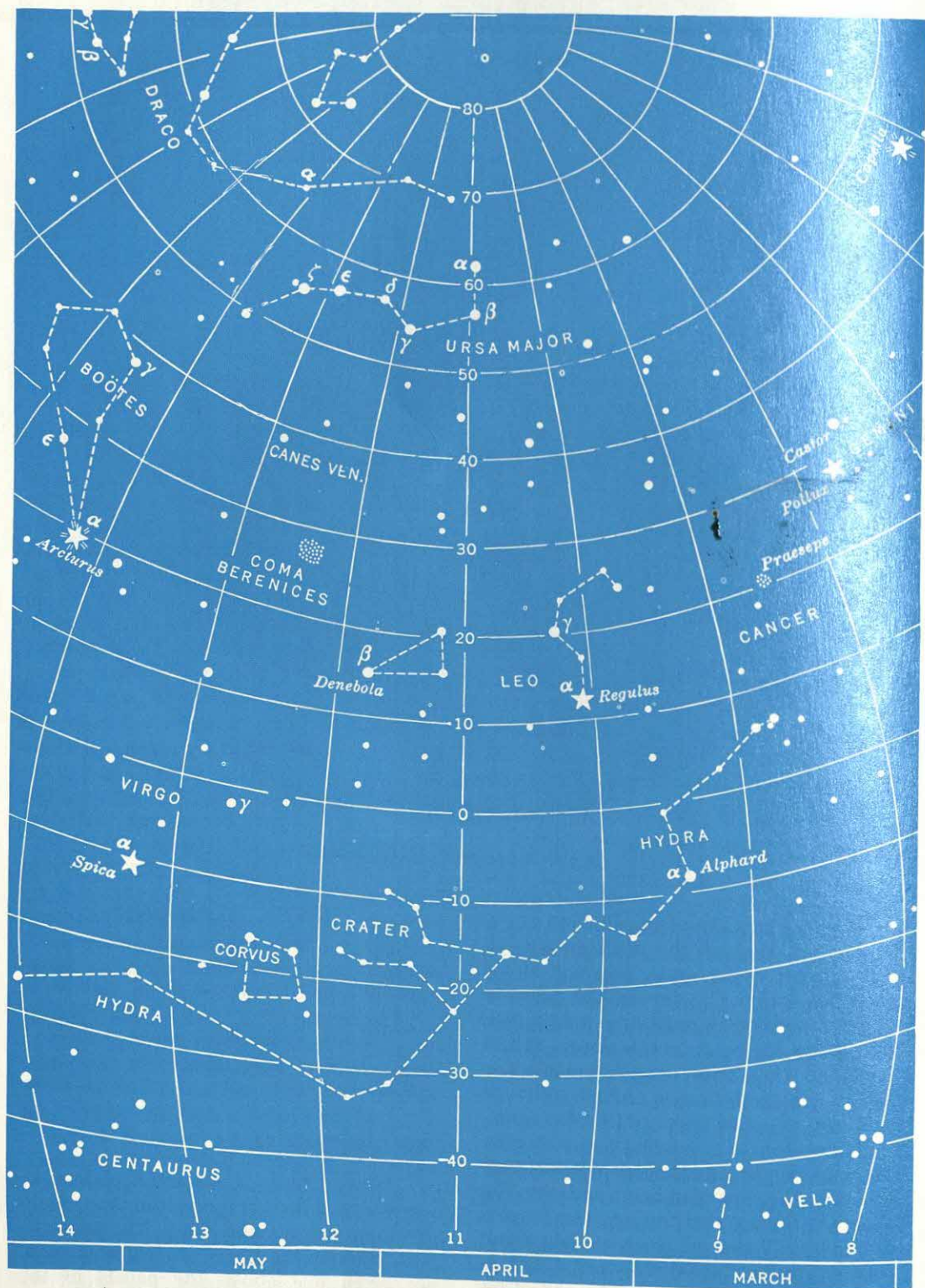
By drawing a line from Epsilon in the Big Dipper, through the polestar and beyond, almost as far again, you will locate the constellation Cassiopeia, the Lady in the Chair. You will recognize it best as a huge **W** or **M**. Next to Cassiopeia is Cepheus, the King. This constellation is not so easy to see as the other groups, although with care you can make out its figure, almost like a tent or a building with a steeple. The long, winding, but inconspicuous, constellation between Cepheus and the Big Dipper is Draco, the Dragon. His head is formed by a **V**-shaped group of stars about halfway around the sky between Cassiopeia and the Dipper's bowl. The end of his tail is marked by faint stars.

GREEK ALPHABET

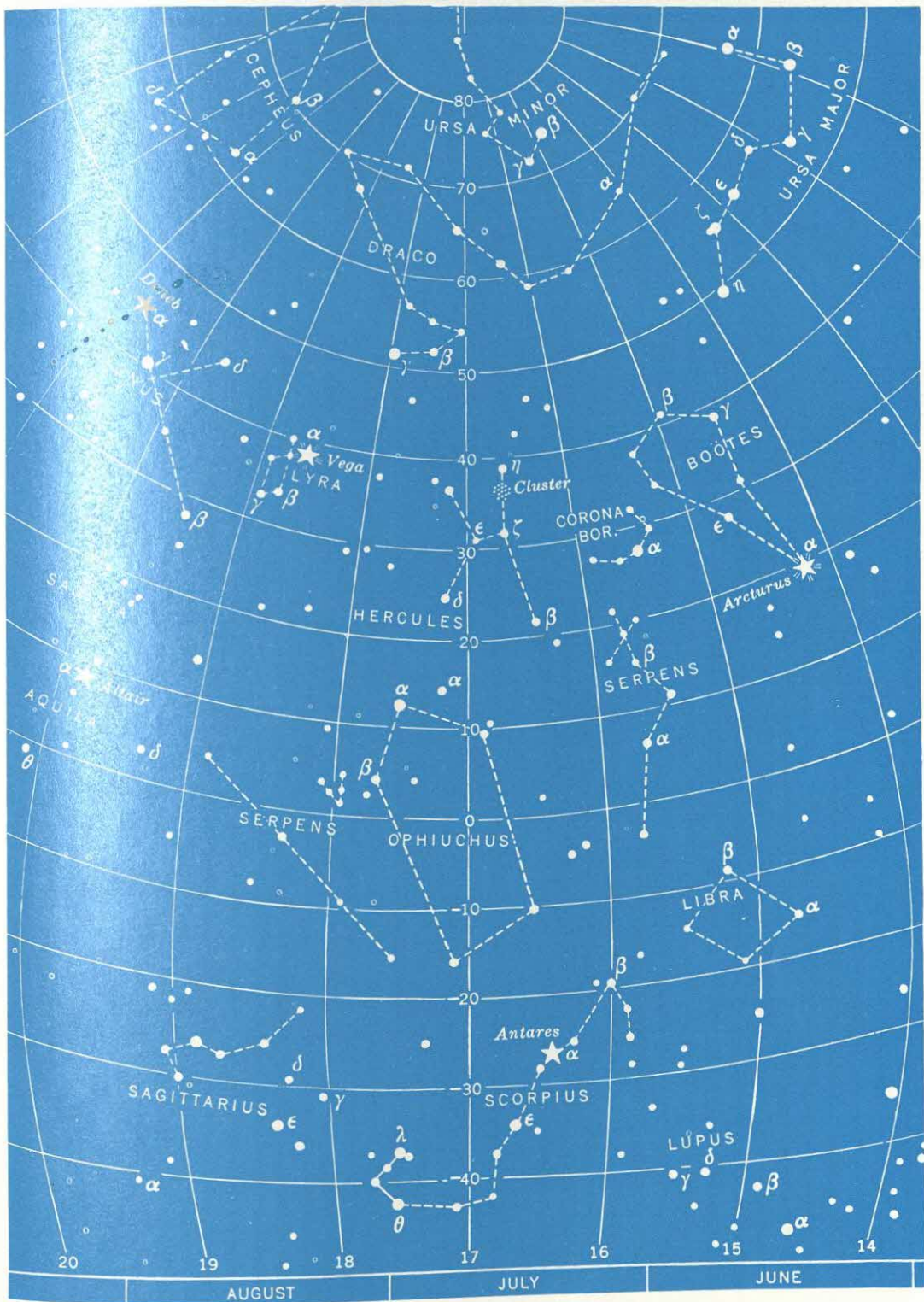
| | | |
|--------------------|--------------------|--------------------|
| α alpha | ι iota | ρ rho |
| β beta | κ kappa | σ sigma |
| γ gamma | λ lambda | τ tau |
| δ delta | μ mu | υ upsilon |
| ϵ epsilon | ν nu | ϕ phi |
| ζ zeta | ξ xi | χ chi |
| η eta | \omicron omicron | ψ psi |
| θ theta | π pi | ω omega |

THE SEASONAL MAPS

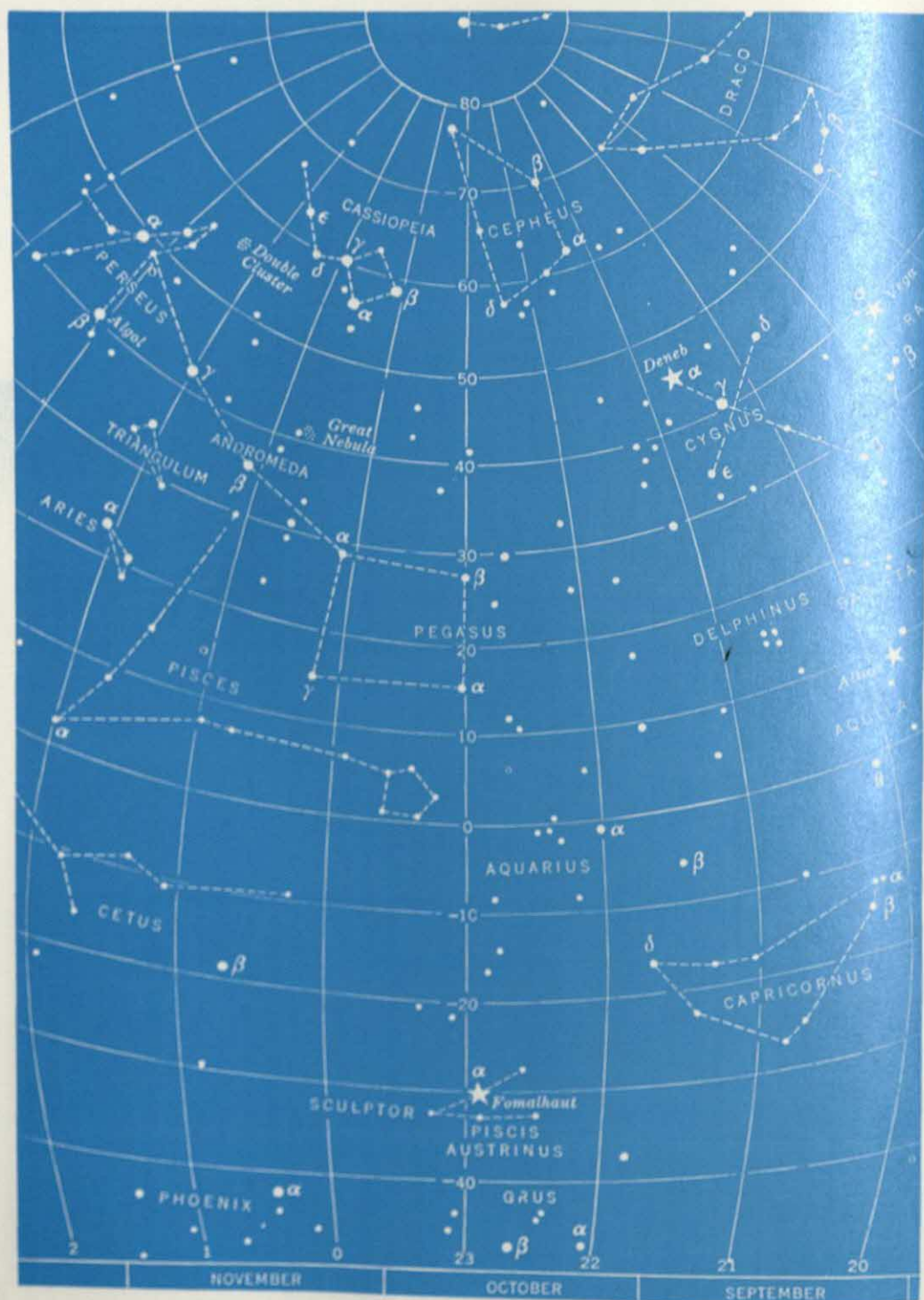
In using Maps 2, 3, 4 and 5, choose the one which shows, at the bottom, the month in which you are observing. The hour circles, marking the hours of right ascension, are given just above the names of the months. Let us imagine that you are looking



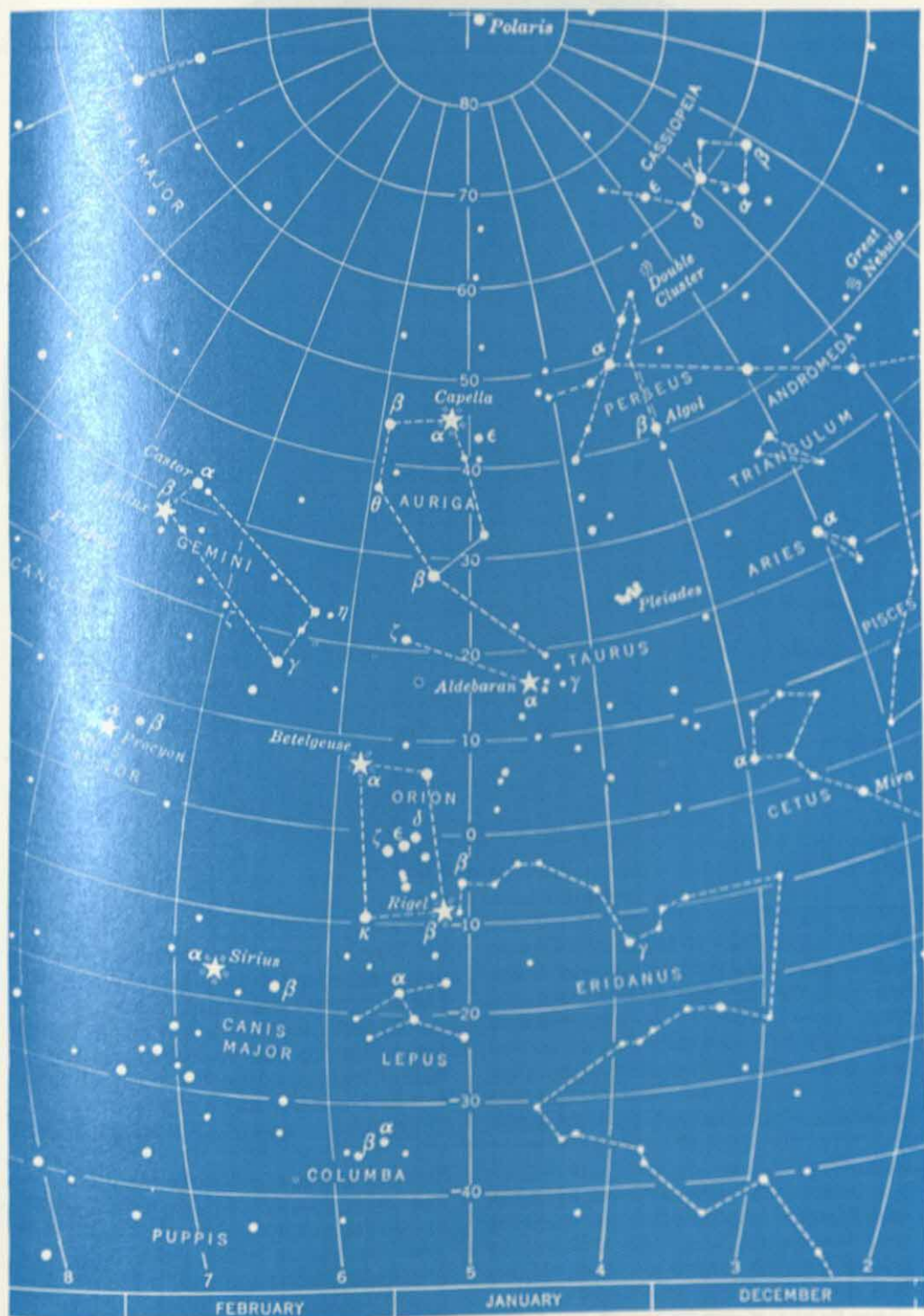
Map 2. Spring constellations, as seen from middle northern latitudes.



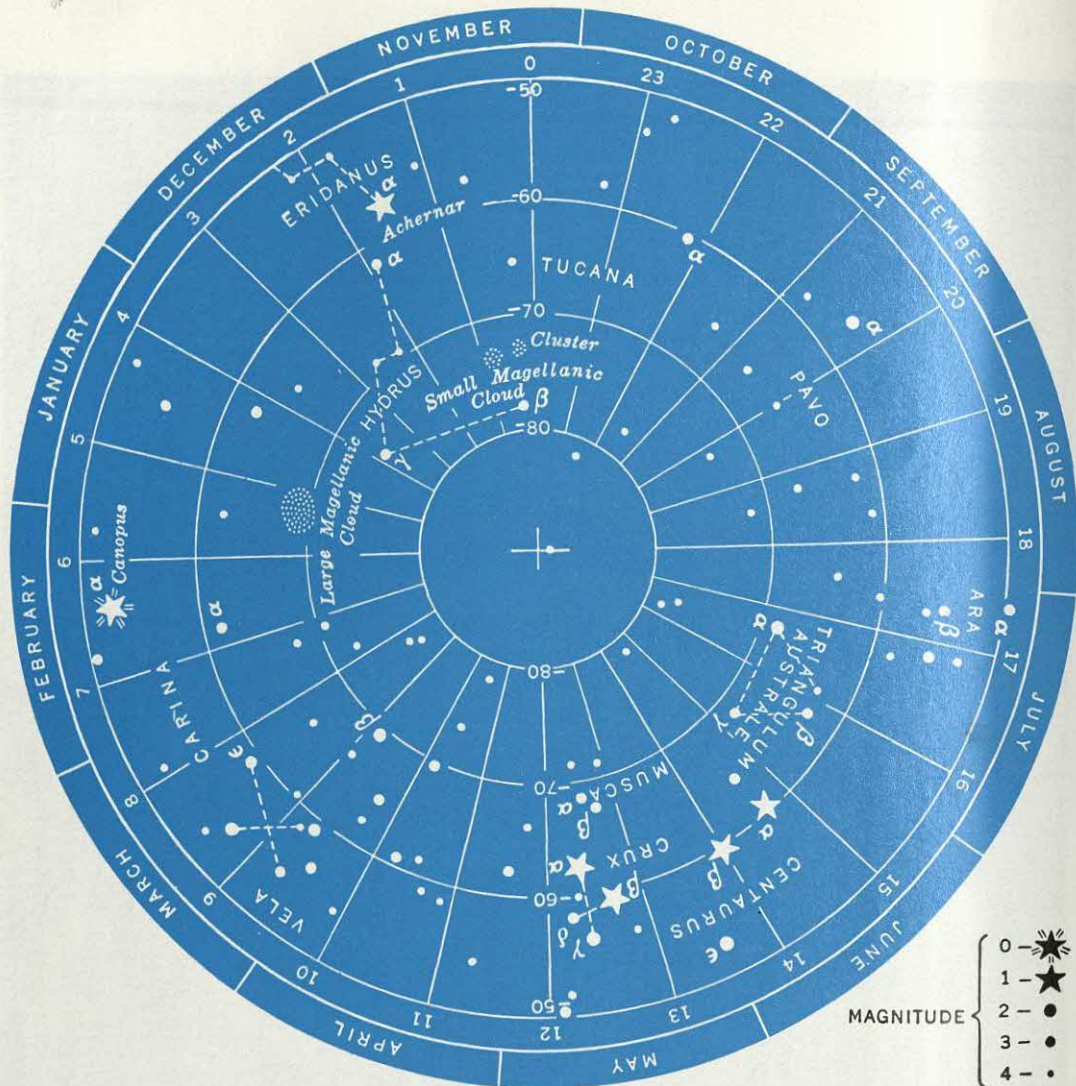
Map 3. Summer constellations, as seen from middle northern latitudes.



Map 4. Autumn constellations, as seen from middle northern latitudes.



Map 5. The winter constellations, as seen from middle northern latitudes.



Map 6. The southern constellations

at the stars on May 20 (Map 2). This date will be about two thirds of the way to the left from the line that divides April and May. Hour circle 13 is approximately above this date. The stars that are on that hour circle will be about on your meridian at 9 P.M. standard time. The ones at the top of the map will be in the north, above Polaris; those at the bottom, in the southern sky.

To locate the stars in the sky when you use Maps 2, 3, 4 and 5, hold the map over your head as you face north, with the top side of the map toward the north. If you face north, you will see the Big Dipper above Polaris; if you face south, the bright star Spica will be almost on your meridian

(on May 20, 9 P.M. standard time). To help you orient yourself, Maps 2, 3, 4 and 5 give half of the north circumpolar stars.

How do you find on your map the stars that will be on your meridian at, say, 11 P.M. or 7 P.M. on May 20? You will recall that at 9 P.M. the meridional stars will be those on hour circle 13. To determine what stars will be on your meridian for later hours of a given night, move ahead (that is, to a higher hour-circle number) by the corresponding number of hours of right ascension. For earlier hours, move back to a lower hour-circle number.

SPRING SKIES

By continuing the curve of the Big

Dipper's handle away from the bowl, you will locate the bright star Arcturus, in the constellation Boötes, the Herdsman. Arcturus is one of the first bright stars that appears over the eastern horizon in the bright spring evenings.

Going back to the Big Dipper, we again continue the curve of the handle to Arcturus and on down to the next brightest star, Spica, in the constellation Virgo, the Maiden. Virgo is one of the twelve constellations of the zodiac, the narrow belt in the sky through which the planets, the sun and the moon appear to move. You will notice that none of the zodiacal constellations are far from the celestial equator, which is marked 0 on the maps.

Near the southern horizon is the small four-sided constellation Corvus, the Crow, which precedes Spica into the sky; its top stars point to Spica. To the west of Arcturus is a splendid, though faint, star cluster known as Coma Berenices—the Hair of Berenice.

Once again we go back to the Big Dipper as a guide group. Extend the line of the pointer stars backward, away from the pole, and they will lead you to Leo, the Lion, another constellation of the zodiac. The part of the Lion to the west resembles a sickle. The other conspicuous group of stars in Leo is a right-angle triangle, to the east of the sickle. Denebola, or Beta, marks the tip of the Lion's tail and Regulus, or Little King, marks the Lion's heart, at the end of the sickle's handle.

To the west of Leo is Cancer, the Crab, also in the zodiac. It is not a well-defined constellation, but if you look out of the corner of your eye you will be able to make out, on a clear night, the faint star cluster, Praesepe, sometimes called the Beehive. Stretching across the southern sky south of Virgo, Corvus, and Leo is the long, faint figure of Hydra, the Sea Serpent. The head is composed of five stars south of Praesepe.

SUMMER SKIES

Map 3 shows the stars that you will see best in the night sky of summer. Just to the east of Boötes is a beautiful crown of stars

which you cannot miss. It is Corona Borealis, the Northern Crown. Its brightest star is Alphecca, sometimes known as Gemma, the gem in the crown. All the other stars of the group are of fourth magnitude.

Just south of Corona Borealis you will notice, if you look carefully, an X made up of five faint stars. It is the head of Serpens, the Serpent. This constellation is closely associated with the large, clearly defined figure of Ophiuchus, the Serpent Bearer of Greek legend. Another group of stars, one to the east of Ophiuchus, is also called Serpens. The group containing the head is known as Serpens Caput (head) and the other as Serpens Cauda (tail). In your imagination you must unite them, and the triangle of stars to the west of Ophiuchus, to form one long serpent stretching across the figure of the Serpent Bearer.

One of the most striking constellations in the heavens is Scorpius, the Scorpion, which is visible in middle northern latitudes just above the southern horizon in the summer. It belongs to the zodiac. To some people it looks a bit like a fishhook. Antares, the brightest star in the constellation, marks the heart of the Scorpion. To the west of Scorpius is Libra, the Scales, another zodiacal constellation.

To the east of Scorpius is Sagittarius, the Archer, also a zodiacal constellation. Some stargazers see in this constellation a teakettle or a little dipper. Through it runs the brightest part of the Milky Way. As we look in this direction, we are gazing toward the center of our galaxy and out toward the opposite rim.

To the east of Corona Borealis is the great roughly H-shaped figure of Hercules. The Alpha of this constellation, Ras Algethi, is not joined in our maps by dotted lines to the rest of the constellation. It is very close to Ras Alhague, the Alpha of Ophiuchus. Between the stars Zeta and Eta is one of the most beautiful star clusters in the whole sky. It can just be glimpsed with the unaided eye.

Following Hercules over the northeastern horizon comes Lyra, the Lyre, a small but extremely beautiful constellation,

composed of a parallelogram and a triangle, joined together. The outstanding star of this group is Vega, which is brighter than first magnitude. It is the brightest star that we can see in northern latitudes in the summer time and the third brightest star in the whole sky. The northernmost star of the triangle of which Vega forms part is Epsilon Lyrae; it is a famous double double star, which you may be able to make out with your unaided eye.

To the east of Lyra is Cygnus, the Swan. It is often called the Northern Cross, though the stars that form the cross are only some of those that make up the Swan. The star at the Swan's tail is Deneb. It is a fine blue star and forms the top of the Cross.

In the Milky Way, not very far from the foot of the Northern Cross, is a first-magnitude star that is the central one of three. The bright star is Altair in the constellation Aquila, the Eagle; the two stars that flank it are much fainter.

AUTUMN SKIES

Map 4 shows us the autumn constellations. Using Altair and its companion stars in Aquila as pointers to the south, we find the rather faint but large group of stars that comprise the zodiacal constellation Capricornus, the Sea Goat. Actually the constellation looks like a tricorn hat upside down.

Between Capricornus and Cygnus are two small, inconspicuous groups called Delphinus, the Dolphin, or Job's Coffin, and Sagitta, the Arrow. East of these groups is Pegasus, best known for its conspicuous great square of stars. The star Alpheratz, in the northeast corner of the Square, forms part of both Pegasus and the constellation Andromeda.

Stretching from the Square's northeastern corner, you will see an almost straight line of three fairly bright stars—Alpha, Beta and Gamma Andromedae. Almost directly above Beta, which is called Mirach, you may be able to make out with the unaided eye a faint, elongated patch of light. This is the Great Nebula in Andromeda—a galaxy of thousands of millions of stars very much like our own Milky Way

galaxy. On Map 4 you will see that the line connecting the three stars of Andromeda has been prolonged almost to Algenib, the Alpha of the constellation Perseus.

Perseus is in the part of the sky from which a famous meteor shower—the Perseids—appears to come between about the 10th and the 12th of August. The Beta of Perseus—Algol—is one of the most fascinating stars that the amateur astronomer can watch. For almost exactly two days and eleven hours, Algol shines steadily as a star of 2.3 magnitude. Then, in a period of about five hours, it decreases to magnitude 3.5. In a second period of five hours it regains its former brilliance and once more remains at that brightness for about fifty-nine hours. This startling change is due to the fact that Algol consists of two stars revolving around their common center of gravity, with the edge of the system turned toward the earth. One star is much brighter than the other. When the fainter star comes between us and the brighter component, Algol seems to be dimmed.

Directly below the Great Square of Pegasus, in Map 4, is a circle of stars that marks the Western Fish, also called the Circlet, in the long ribbonlike constellation Pisces, the Fishes. To the east of Pisces is the constellation Aries. To the west there is a veritable cascade of stars—the constellation Aquarius, the Water Carrier, in the zodiac. None of the stars of Aquarius is bright.

Below Aquarius and close to the southern horizon is the bright star Fomalhaut in Piscis Austrinus, the Southern Fish.

WINTER SKIES

There is no more superb celestial sight than the sky of winter, shown in Map 5. One of the most splendid of all the constellations is Orion, the Giant Hunter of the heavens. As you can see from the map, Orion is on the meridian about 9 P.M. on February 1, about 11 P.M. on January 1 and 1 P.M. on December 1. The figure of Orion is roughly rectangular and easy to find. His right shoulder (he is facing you) is marked by the first-magnitude star Betelgeuse, a gigantic red sun hundreds of times as large

as our own sun. Rigel, a gorgeous blue-white star, marks his left foot. Dividing Orion's rectangle in two are the three stars of the belt, from which hangs the Giant's sword. The top star of the belt, Delta, is almost exactly on the celestial equator. One of the most interesting objects in this constellation is the great gaseous nebula—M42—which surrounds the middle star of the sword. You can merely glimpse it with the unaided eye.

South of Aries we find the head of the constellation Cetus. The head is made up of five stars. Cetus has a remarkable long-period variable star Mira, known as the Wonderful. When Mira is at its most brilliant it is usually about magnitude 3.5, although it may be even brighter. At its faintest it is about ninth magnitude, far below the level of naked-eye vision. Mira goes through all its changes in a period of about 330 days.

Northwest of Orion we come upon the first-magnitude star Aldebaran, which represents the eye of Taurus, the Bull, a constellation of the zodiac. The face of the Bull is a V-shaped group of stars—an open star cluster called the Hyades. Taurus boasts another even more beautiful open cluster of stars, the Pleiades, to the northwest of the Bull's face. The Pleiades mark the shoulder of the Bull. (Only the forepart of the Bull is represented.) Most people can make out six of the Pleiades with the unaided eye. Beta Tauri, known as El Nath, "that which butts," marks the tip of the left horn. It also belongs to the five-sided constellation Auriga, the Charioteer. Its brightest star is Capella, the She-goat. Near Capella is a small triangle of three stars representing Capella's kids, a very good guide group to help you make sure you have actually found Capella.

By extending the line of Orion's belt to the southeast, we discover Sirius, the Dog Star, and the brightest star in the entire sky. It is in the constellation Canis Major, the Greater Dog. Forming an almost equilateral triangle with Betelgeuse and Sirius is Procyon, another first-magnitude star, in Canis Minor, the Lesser Dog.

At Orion's left foot, Rigel, the long,

meandering constellation Eridanus begins. Most of its stars are faint.

Gemini, the Twins, a zodiacal constellation, is easily recognized by its two parallel lines of stars, with Pollux (Beta) at the northern head of one line and Castor (Alpha) at the northern end of the other. The constellation can be located by drawing a line from Rigel, in Orion, through Betelgeuse and as far again beyond.

SOUTHERN CONSTELLATIONS

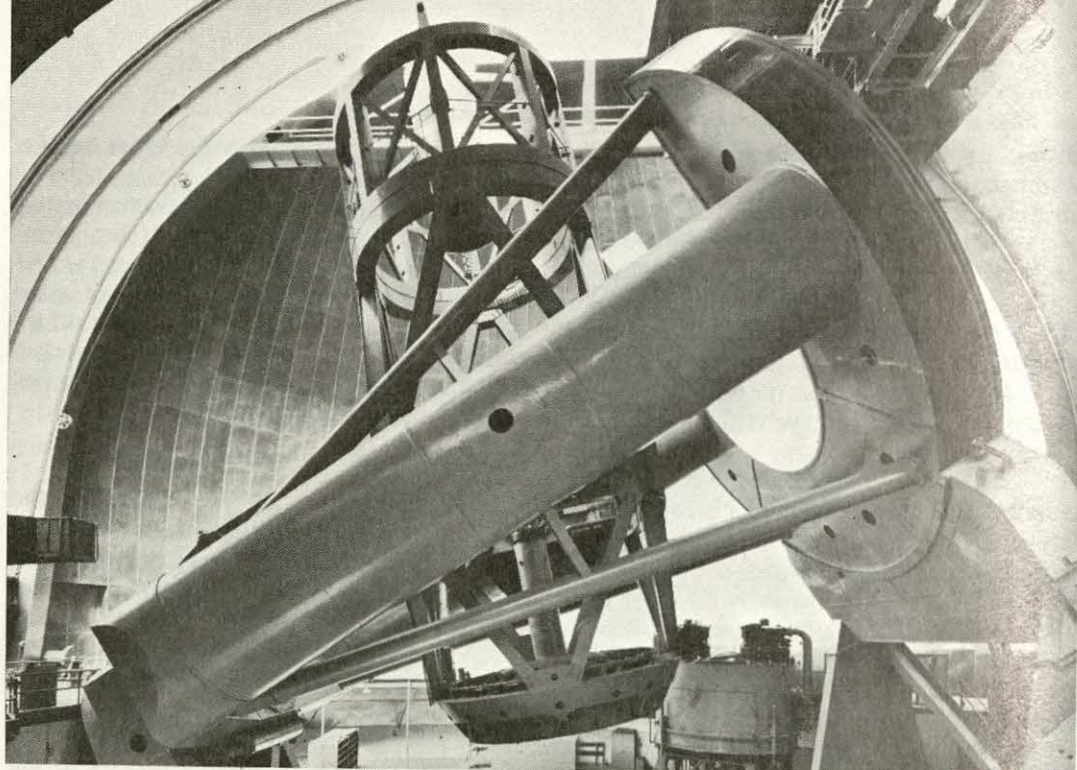
We turn now to Map 6, showing the circumpolar stars visible from midsouthern latitudes. The south celestial pole is at the center of the cross in the middle of the diagram, but there is no bright star to mark it conveniently for navigators in southern seas or skies. There is a guide to the south celestial pole, however: the constellation Crux, the Southern Cross. The longer axis of the Cross points almost directly to the south pole of the heavens.

The two bright stars Alpha and Beta Centauri point to the top of the Southern Cross—the star Gamma Crucis. Alpha Centauri is a double star. It is generally held to be the star nearest to the earth (except for the sun); it is about 40 million-million kilometers away. However, some astronomers think that its faint neighbor, Proxima Centauri, may be a bit closer. Close to the Southern Cross is a famous dark nebula, the Coal Sack.

The Magellanic Clouds, clearly marked on Map 6, are hazy objects which, through a telescope, are resolved into masses of stars, nebulas and star clusters.

About the first of December, at the end of the constellation Eridanus, the star Achernar is seen above the south pole. It is especially arresting since it is the one brilliant object in that long stream of stars. It forms, roughly, a right-angle triangle with the two Magellanic Clouds.

Canopus, second only to Sirius among the stars in brightness, is in the constellation Carina, the Keel. This constellation was once part of the big constellation Argo Navis, the Ship Argo, which in modern times has been broken up into several smaller constellations.



Mount Wilson and Palomar Observatories
The 500-centimeter reflecting telescope in position at the Hale Observatory on Palomar Mountain in California.

TELESCOPES

by Laurence W. Fredrick III

Have you used field glasses or binoculars when observing a sports event or trying to identify a small bird high in a tree? If you have, then you know that after having focused the binoculars you were able to see the playing field or the bird more clearly. Did you realize that you were focusing and using a pair of small telescopes? Maybe not. You and many people may think of telescopes only as instruments for observing the sky. And, indeed, this is the main use of the telescope. Long before the beginning of the space age, people used telescopes to go exploring among the stars and planets in the sky. And even today, with space satellites and manned space flights, the telescope is still the basic tool of the astronomer.

We are not certain who built the first simple telescope. Some historians give the credit to a Dutch lens maker, Hans Lipperhey, who lived at the beginning of the 17th

century. The first person to use a telescope to look up at the sky, however, was the Italian scientist Galileo. In 1609 he became the first person to see the craters and mountains of the moon, the satellites of Jupiter, and the phases of the planet Venus. His work marked the beginning of modern observational astronomy.

The telescopes with which we are most familiar are optical telescopes. Optical telescopes are used to observe visible light. There are other kinds of telescopes as well, designed to observe other forms of radiation: radio waves, infrared and ultraviolet light, X rays, and so on. The general principles by which telescopes work are just about the same in all cases, but there are great practical differences in design.

WHAT TELESCOPES DO

The word "telescope" means "to see at a distance." This is the purpose for

which Galileo and other early astronomers used the instrument and for which smaller telescopes are used on the earth today. However, the main purpose of the giant modern telescopes is not to see fine details at a distance. Instead, it is to collect light.

The larger the telescope, the more light it will collect. To understand this, imagine a gentle rain falling on two buckets of equal size. One bucket has a cover over it, with a hole one half the diameter of the bucket. The other bucket has no cover. Clearly, the open bucket will collect more rain than the half-covered one. It will do so in proportion to the difference in the size of the bucket openings. Similarly, a telescope with a larger lens opening will collect more light than a telescope with a smaller opening. The great value of the giant telescopes is that they can collect light from very faint objects in the night sky.

The other important purpose of a telescope is to let us see the fine details of distant objects. The ability of the telescope to reveal such details is known as its *resolving power*. The resolving power of a telescope is directly proportional to the diameter of the telescope's lens or mirror. Thus it would seem that the largest telescopes should have the greatest resolving power.

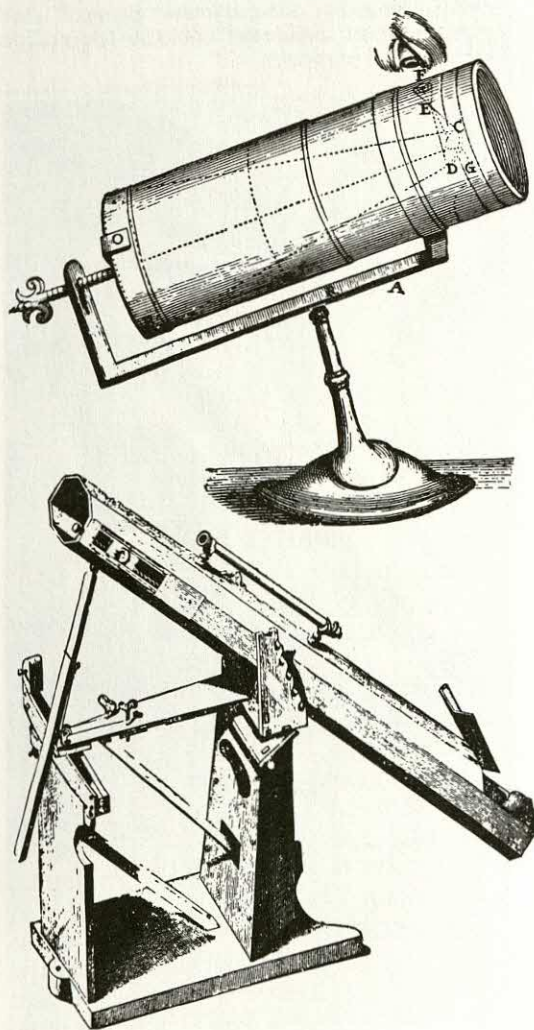
However, things do not work out that way for earth-based telescopes. The reason is the earth's unsteady atmosphere. The atmosphere has regions of different density. As the atmosphere moves about, these differences cause light rays to bend this way and that. This bending of light rays interferes with the resolving power of large earth-based telescopes. In fact, the earth's atmosphere is so unsteady that the resolving power of a 25-centimeter telescope is as good as an astronomer is going to achieve. Large telescopes do not have any better resolving power.

LOCATING A TELESCOPE

A telescope must be properly located if it is to collect as much light as possible. A dark sky as free as possible from human lights is needed to observe faint celestial objects. Also, to get a steady image, the atmosphere should be clear and dry. These

conditions are usually best met on remote mountains. Unfortunately, mountains that are remote from civilization one year may not be so far away a few years later. Sometimes compromise locations are chosen to provide better coverage for one or the other telescopes at different times of the year. For example, telescopes located in California have their best observing times in the spring and summer, while those located in Arizona have theirs in the fall and winter.

Simple reflecting telescopes were first built in the 17th century. Below: one of Sir Isaac Newton's telescopes. Bottom: one invented by the English inventor John Hadley.



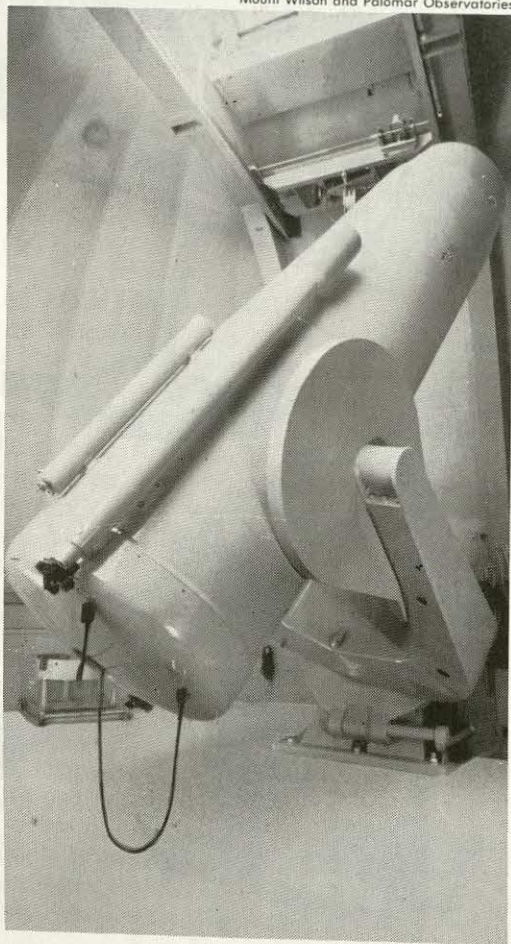
REFRACTING AND REFLECTING TELESCOPES

There are three general types of optical telescopes: refracting telescopes, or refractors; reflecting telescopes, or reflectors; and telescopes that combine features of both types.

Refracting telescopes, the type often depicted in comic strips, are easy to understand. A refracting telescope consists of a long tube. At one end of the tube is a large convex (outward-bulging) lens. This lens is called the *objective*. It is the part of the telescope that gathers light from a celestial object. The lens bends the light rays. This

Palomar Mountain 120-centimeter Schmidt telescope. Schmidt telescopes combine features of reflectors and refractors.

Mount Wilson and Palomar Observatories



bending is called *refraction*. It brings the incoming light rays to a focus to produce an *image*. The image is then magnified, or made to seem larger, and viewed through a lens device known as an *eyepiece*. The eyepiece is located at the other end of the telescope tube.

Reflecting telescopes have a more complex design. In a reflecting telescope, a glass mirror with a curved surface shaped like the headlamp of an automobile is used as the objective. The mirror gathers and focuses incoming light rays to a point called the *prime focus*. This point is in front of the mirror. How, then, do you manage to see the image without getting in the way of the incoming light? The fact is that in all but the largest telescopes the prime focus cannot be reached by the observer. A second mirror is placed near the prime focus and is used to relay the light beam elsewhere—to a place convenient for the observer. In some cases, a flat mirror, set at a 45° angle to the axis of the telescope tube, is used to place the focus just outside the tube. This arrangement is known as *Newtonian focus*. This design is named for the 17th century mathematician-astronomer Isaac Newton, who first developed it. This type of focusing arrangement is very popular in small, amateur reflecting telescopes.

Another solution involves using a curved mirror to intercept the light beam. This mirror sends the beam back through a hole in the main mirror so that it comes to a focus behind that mirror. This arrangement is known as *Cassegrain focus*. It is the one probably most used for reflecting telescopes by professional astronomers. There are other possible focusing arrangements, however.

The common caricature of an astronomer peering into the eyepiece of a telescope is really not accurate. Astronomers rarely spend any time doing this. Most of the time that a telescope is being used, it is guided and moved automatically. Desired images of the sky are recorded on photographic plates. When it is necessary for the astronomer actually to “see” something through a telescope, it is usually done with a television system.

COMBINATION TELESCOPES

The third type of optical telescope is one that combines refracting and reflecting methods. Such telescopes are often used to achieve a wide field of view or to achieve a very large-scale image in a short distance.

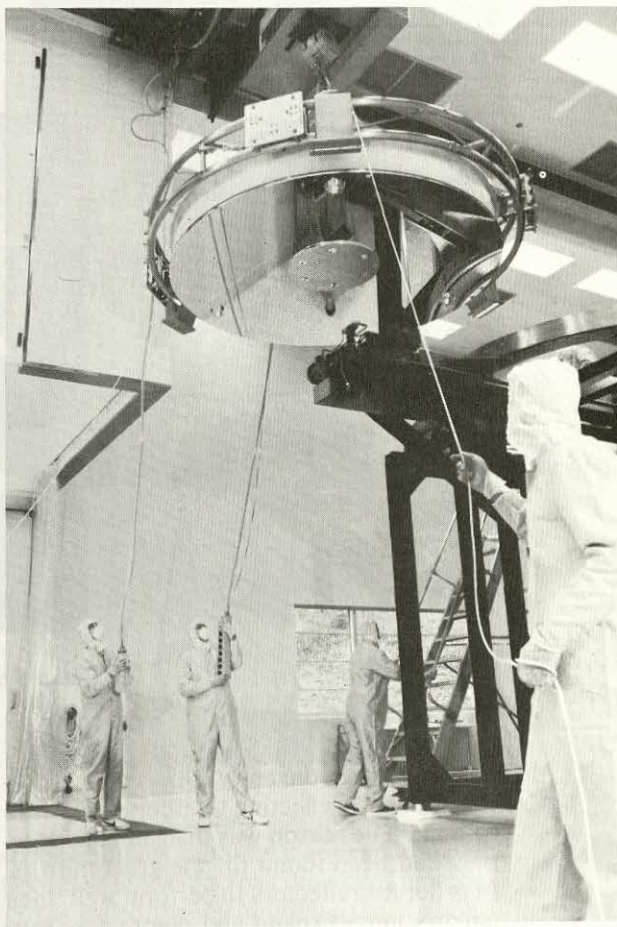
Light that enters such a telescope passes through a very weak lens called a *corrector plate*. It then strikes a spherical mirror and comes to a focus. Unlike parabolic mirrors, spherical mirrors by themselves do not make good telescope objectives. They do not bring all of the light to the same focus. That is why the weak corrector plate is used—to correct this unwanted condition.

The first and most famous telescope of this combined type is called the Schmidt telescope—after the German optician Bernard Schmidt, who designed and built it. The most common telescope for providing large-scale images is the Bouwers-Maksutov system, named for the Dutch and Russian opticians who developed the system independently of each other. In this system, the surfaces of the corrector plate are also spherical, as is the mirror. The system is used as a telephoto lens in photography.

LENSES VERSUS MIRRORS

The very first telescopes were simple lenses, but the advantages of mirrors were quickly recognized. Mirrors can be supported at the back as well as at the sides, making larger systems easier to design. Also, the glass of the mirror does not need to be perfect except at its reflecting surface, since the light does not pass through it. The glass of a lens has to be free of faults.

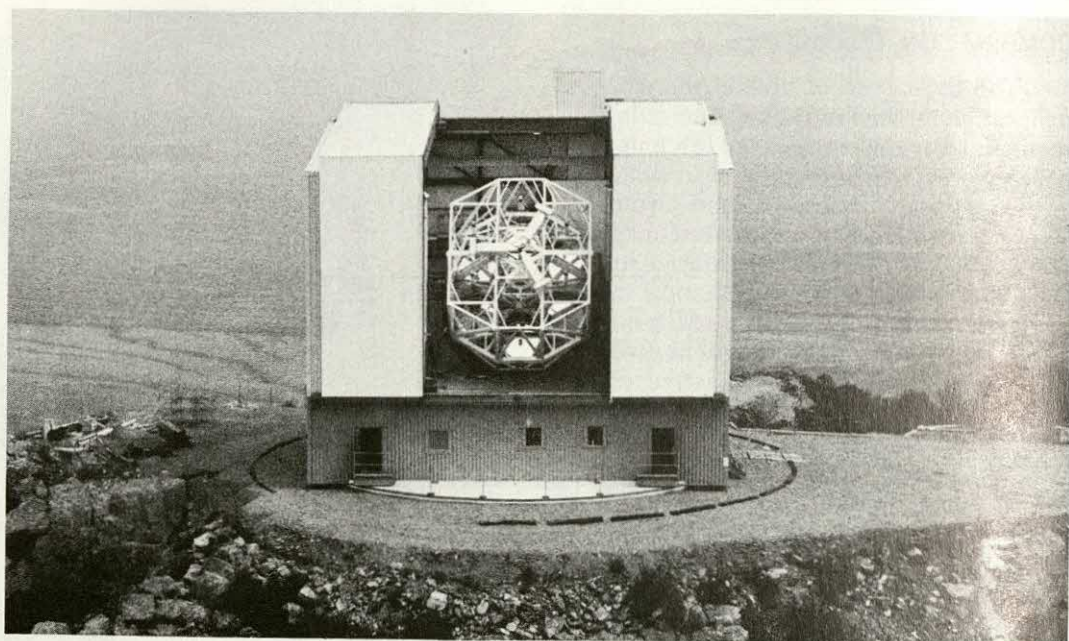
Actually, the earlier large telescopes were reflectors that used not glass but metal mirrors, made of speculum—a nickel compound that took a good polish. The speculum tarnished easily, however, and it was a nuisance to take the telescope apart every few weeks to repolish the mirror. In addition, the speculum performed poorly, producing very distorted images when temperature changes occurred. Therefore these telescopes fell into disfavor. Until the craft of making large mirrors developed further, very large reflectors would not be built.



Courtesy of The Perkin-Elmer Corporation

The 240-centimeter primary mirror for NASA's space telescope is lifted from the vacuum chamber in which it was given its reflective coating.

On the other hand, a major problem of early refracting telescopes was that the lens distorted the image. A lens bends light of different wavelengths at different angles. This results in a spread of color known as *chromatic aberration*. Toward the end of the 17th century, astronomers learned how to combine two or more lenses to correct for chromatic aberration. Thereafter, refractors became the most commonly built telescopes over the next two centuries. The largest of all is the 102-centimeter telescope at Yerkes Observatory in Wisconsin, U.S. Lenses can hardly be made larger than that, because glass at this size begins to deform under its own weight.



University of Arizona

The Multiple-Mirror Telescope uses six mirrors to create the light-gathering power of a large telescope.

Starting in the early 20th century, larger and larger reflecting telescopes began to be built. The reason was that suitable materials had been found for the great mirrors. In order for reflectors to perform well, their mirrors must expand and contract as little as possible in response to temperature changes. Glass and metal will not do well, but astronomers found that Pyrex and quartz are good for this purpose. Today there are materials available that scarcely change at all with temperature. None of the great modern reflecting telescopes has a mirror made of common glass.

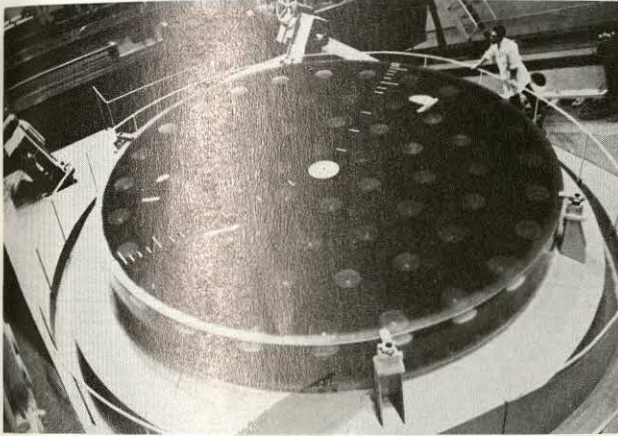
One of the solutions to the problem of building ever larger telescopes to reveal ever more distant objects has been the development of the Multiple-Mirror Telescope (MMT). On the summit of Mount Hopkins in Arizona, an array of six computer-controlled, 183-centimeter primary mirrors work in concert to produce an image that under ideal conditions is equivalent to that of a reflector with a diameter of almost 700 centimeters. A joint project of the Smithsonian Astrophysical Observa-

tory and the University of Arizona, the MMT was inaugurated in 1979. It is housed in a four-story, cubelike observatory that rotates smoothly with the telescope itself.

TELESCOPE MOUNTINGS

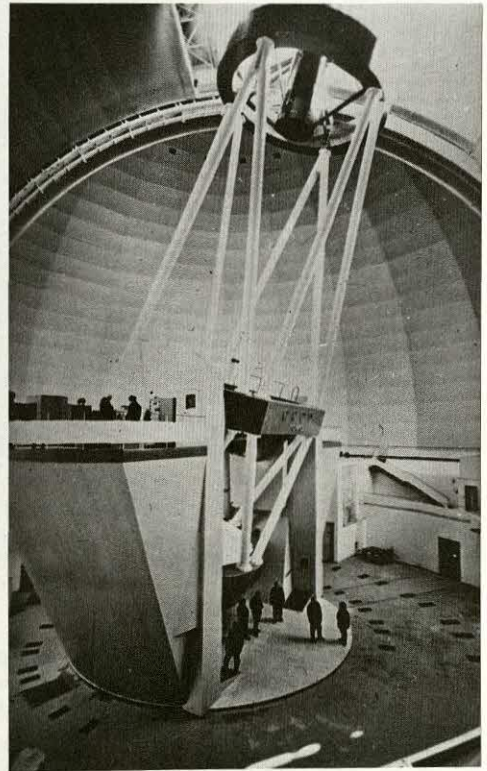
Most telescopes are not fixed rigidly in one position. They are mounted so that they can be turned in one direction or another. The reason for this is that celestial objects appear to change their positions in the sky. If light is to be collected from a faint star, the telescope has to be able to follow the star. Most of the movement we observe is due to the fact that the earth is rotating on its axis. As the earth rotates from west to east, celestial objects appear to rise in the east and move westward.

The earliest telescopes were simply moved by hand. For a long while thereafter, telescopes were driven by weights, just as cuckoo clocks are. Today almost all telescopes have motor drives that are controlled electronically. Many have automatic devices, and some use computers, to relieve the astronomer of having to be con-



both photos, Novosil, from Sovfoto

The largest reflecting telescope in the world has been constructed by the Soviets in the Caucasus Mountains. The 600-centimeter parabolic mirror is shown above at its final polishing before it was installed in the telescope. The telescope has an altitude-azimuthal mounting. It can swing up and down and has a rotating base (right).



cerned with the details of telescope setting and movement. The telescopes are very carefully balanced on their mountings so that they can move smoothly and with little effort.

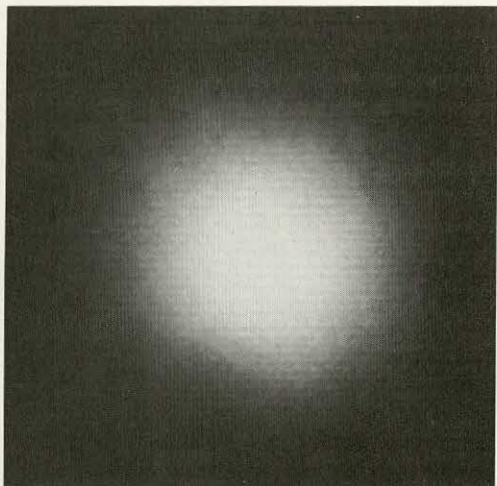
Astronomers refer to the moving of a telescope as *tracking*, since the instruments are actually tracking objects as they move across the sky. The kind of tracking that is possible depends on the way in which the telescope is mounted. Optical telescopes (and many radio telescopes) are mounted *equatorially*. That is, they are mounted in the equatorial coordinate system used in mapping the stars. This coordinate system corresponds to the system of latitude and longitude used to map the earth's surface. In the equatorial system, latitude is called *declination* and longitude is called *right ascension*.

An equatorial mounting has two axes around which the telescope may be turned. One is the *polar axis*, which allows the tele-

scope to follow a star as it appears to move from east to west across the sky. The other axis is called the *declination axis*. It allows the telescope to be pointed toward stars higher or lower above the horizon.

Suppose a given star is to be tracked. In theory, an equatorially mounted telescope need only move around its polar axis as it follows the star across the sky. This is true for small telescopes. However, large telescopes must use the declination axis as well. The reason is that the atmosphere acts as a lens, bending the light from the star and causing it to appear slightly higher in the sky than it actually is. This effect becomes stronger the nearer the star is to the horizon. Therefore the telescope must keep moving slightly around its declination axis to adjust for this effect and keep the star in focus.

Not all telescopes are mounted in this way. For example, there is a special-purpose telescope known as a *photographic*



Kitt Peak National Observatory

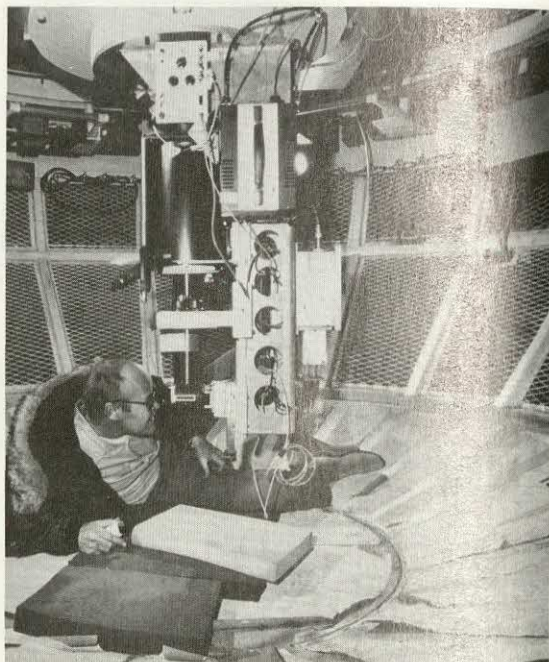
Above: reconstructed image of the star Betelgeuse. Astronomers at the Kitt Peak Observatory were the first to observe the surface of a star other than the sun. Right: Dr. Roger Lynds, one of the scientists, at the secondary focus of the 400-centimeter reflecting telescope used in the work.

zenith tube. This is a refracting telescope that is always pointed toward the zenith—the point in the sky directly above the observer. A photographic plate holder is located directly behind the telescope lens, facing downward. As a star passes near or through the zenith, its image is reflected off the surface of a pool of mercury below the lens and onto the photographic plate. This kind of telescope is used for keeping an accurate check on celestial time.

COLLECTING OTHER RADIATION

Modern astronomy is interested in other forms of radiation from the sky in addition to visible light—radio waves, infrared and ultraviolet radiation, and so on. The telescopes used for collecting such radiation must be somewhat different in design from visible-light telescopes.

These design differences involve resolving power. The resolving power of a telescope is proportional to the wavelength of the incoming radiation. All other things being equal, a telescope working at a visible-light wavelength of about 5,000 Ångstroms has twice the resolving power that it does when working at a wavelength of 10,000 Ångstroms—the near infrared. (1 Ångstrom = 0.00000001 centimeter.)



Kitt Peak National Observatory

Opticians polish a telescope mirror to an accuracy of one-eighth the wavelength at which the telescope is to operate. Therefore if a telescope is being made to operate at 200,000 Ångstroms—the far infrared—it needs to be polished to an accuracy only one-fortieth that of the telescope designed to collect visible light at about 5,000 Ångstroms. Such roughly polished telescopes, which are made for collecting in the far infrared, are often called *light buckets*.

The resolving power of a telescope is also inversely proportional to the width of its aperture—its lens or mirror or other receiving surface. That is, at a given wavelength, a telescope with an aperture twice as wide as that of another telescope has twice the resolving power. If a telescope operating at 10,000 Ångstroms is to have the same resolving power as a telescope operating at 5,000 Ångstroms, it must have an aperture twice as wide.

By the time that wavelengths as long as those of radio waves—much longer than the infrared—are reached, this becomes a major problem. A radio telescope working at a wavelength of 50 centimeters would need to have an aperture 1,000,000 times



Lick Observatory

The 300-centimeter reflecting telescope at the Lick Observatory at Mount Hamilton, California. It is used to study interstellar space and stars.

as wide as a telescope working at a wavelength of 5,000 Ångströms, if the same resolving power were to be provided. This would amount to a width of 5,000 kilometers. Such a telescope is clearly impossible. Something else can be done, however. Two radio telescopes placed at a distance from each other but connected by a cable can be made to act as a single telescope. The effect is to achieve a telescope as wide as the distance between the two receivers. Of course, the two smaller telescopes do not collect as much radio energy from the sky as would a single enormous radio telescope. A system of two or more telescopes arranged for this purpose is called an *interferometer*. For a more detailed discussion of radio telescopes and the work done with them—alone and together with optical telescopes—see the article “Radio Astronomy” in *The New Book of Popular Science*.

But what about wavelengths that are very short, much shorter than visible and ultraviolet light? At such short wavelengths the astronomer is dealing with particles. In this case, he often uses particle counters instead of trying to obtain a telescope image.

TELESCOPES OF TOMORROW

By using the various kinds of telescopes that have been described, astronomers are slowly putting together a more and more complete picture of the universe. Telescopes in the near future are not likely to be very different from those of today. There will be small changes brought about by advancing technology rather than revolutionary changes in design and concept. An exciting advance is expected to come in space astronomy, with large telescopes placed in orbit and maintained by space shuttle—free of the earth's obscuring atmosphere. A 240-centimeter reflecting telescope is scheduled for launch by NASA in 1986. As more such instruments are developed, we should gain a much better understanding of the universe.

Balloon-borne telescopes function free of some of the interference caused by the earth's atmosphere. Photo below shows an L-shaped telescope about to be launched.

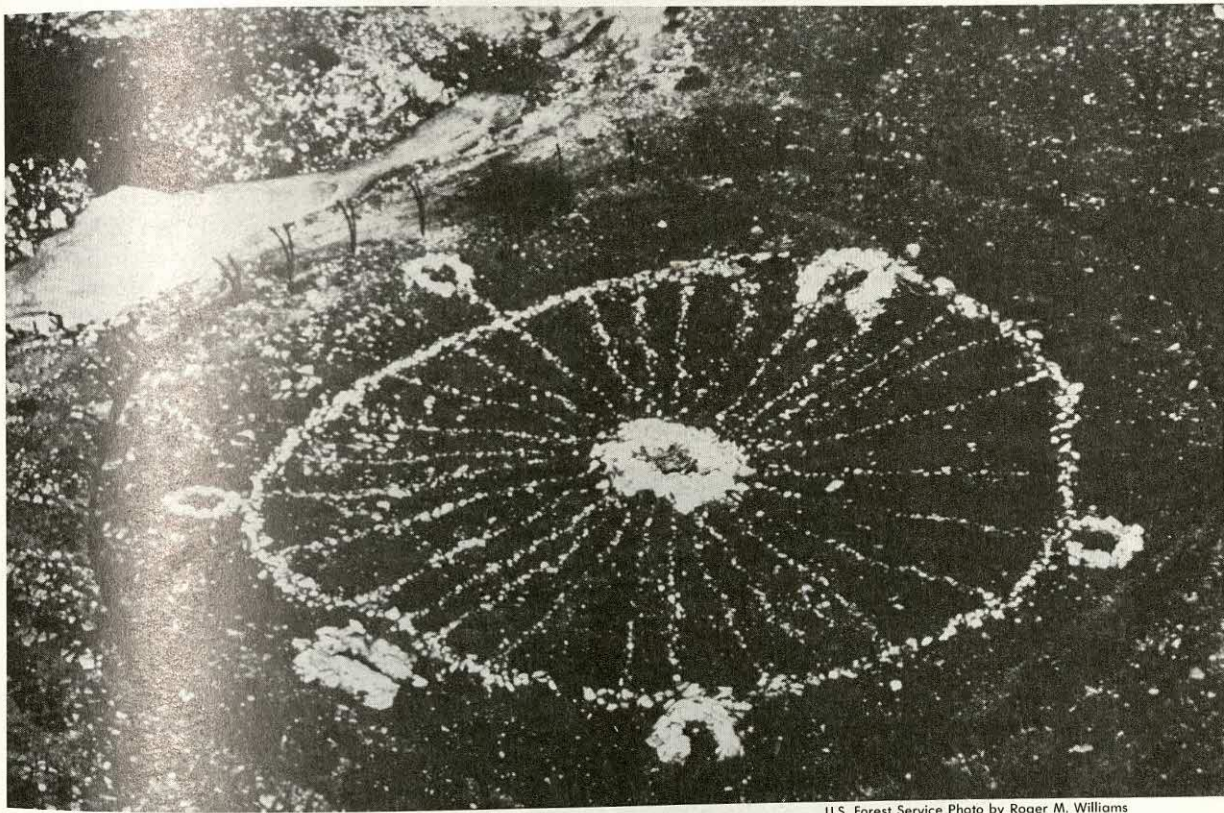
N.S.F. Photo



IMPORTANT OPTICAL OBSERVATORIES AND THEIR TELESCOPES

| Observatory | Location | Type of Telescope | Aperture |
|--|---|-------------------------------------|----------------------------|
| Allegheny O. | Pittsburgh, Pennsylvania | Refracting | 76 cm |
| Argentina National O. | Bosque Alegre, Argentina | Reflecting | 160 cm |
| Astrophysikalisches O. | Potsdam, GDR | Refracting | 80 cm |
| Australian National O. | Siding Spring Mtn., Australia | Reflecting | 380 cm |
| Boyden Station (Armagh O., Dunsink O., Hamburg O., Stockholm O., Uccle O.) | Mazelspoort, Rep. of South Africa | Reflecting Schmidt | 150 cm 81 cm |
| Cerro Tololo Inter-American O. | Cerro Tololo, Chile | Reflecting Schmidt | 400 cm 61 cm |
| Canada-France-Hawaii | Mauna Kea, Hawaii | Reflecting | 360 cm |
| Crimean Astrophysical O. | Central Crimea, USSR | Reflecting | 260 cm |
| David Dunlap O. (U. of Toronto) | Richmond Hill, Ont., Canada | Reflecting | 188 cm |
| Dominion Astrophysical O. | Victoria, B.C., Canada | Reflecting | 185 cm |
| European Southern O. (Belgium, Denmark, France, Holland, Sweden, West Germany) | Cerro La Silla, Chile | Reflecting Reflecting Schmidt | 366 cm 150 cm 100 cm |
| Hale O. (Mt. Wilson and Palomar Observatories) | Mt. Wilson, Palomar Mountain, California | Reflecting Reflecting Schmidt | 500 cm 254 cm 120 cm |
| Harvard College O., Agassiz Sta. | Harvard, Massachusetts | Reflecting | 150 cm |
| Haute-Provence O. | Saint Michel-l'Observatoire | Reflecting | 200 cm |
| Kitt Peak National O. | Kitt Peak, Arizona | Reflecting | 400 cm |
| La Plata O. | La Plata, Argentina | Reflecting | 220 cm |
| Las Campanas O. (Carnegie Institution of Washington) | Las Campanas, Chile | Reflecting | 260 cm |
| Lick O. (U. of California) | Mt. Hamilton, California | Reflecting Refracting | 300 cm 90 cm |
| Lowell O. | Flagstaff, Arizona | Reflecting | 180 cm |
| Mauna Kea (U. of Hawaii) | Mauna Kea, Hawaii | Reflecting | 220 cm |
| McDonald O. (U. of Texas and U. of Chicago) | Mt. Locke, Texas | Reflecting Reflecting | 270 cm 210 cm |
| Mt. Hopkins O. | Amado, Arizona | Reflecting | 690 cm* |
| Mt. Stromlo O. | Canberra, Australia | Reflecting | 190 cm |
| National Astronomical O. of Mexico | Baja California, Mexico | Reflecting Schmidt | 150 cm 66 cm |
| Observatoire de Paris | Meudon, France | Refracting | 83 cm |
| Okayama Astrophysical Station | Kamogata, Japan | Reflecting | 190 cm |
| Ondrejov O. | Ondrejov, Czechoslovakia | Reflecting | 200 cm |
| Radcliffe O. | Pretoria, Rep. of South Africa | Reflecting | 190 cm |
| Republic O. Annexe | Hartbeesport, Rep. of South Africa | Reflecting | 190 cm |
| Royal Greenwich O. | Herstmonceux, England | Refracting | 71 cm |
| Shemakha Astrophysical O. | Shemakha, Azerbaijan | Reflecting | 200 cm |
| Smithsonian Astrophysical O. | Mt. Hopkins, Arizona | Reflecting | 150 cm |
| Steward O. (U. of Arizona) | Kitt Peak, Arizona | Reflecting | 230 cm |
| Tautenburg O. | Tautenburg, GDR | Schmidt | 130 cm |
| South African O. | Johannesburg, Rep. of South Africa | Refracting | 67 cm |
| U.S. Naval O. | Flagstaff, Arizona Washington, D.C. | Reflecting Refracting | 150 cm 66 cm |
| Universitäts-Sternwarte | Vienna, Austria | Refracting | 67 cm |
| University of Padua O. | Asiago, Italy | Reflecting | 180 cm |
| Nice O. | Nice, France | Refracting | 76 cm |
| University of Toulouse O. | Pic du Midi, France | Refracting Refracting | 60 cm 105 cm |
| Uppsala O. (Kvistaberg Station) | Bro, Sweden | Schmidt | 99 cm |
| Yerkes O. (U. of Chicago) | Williams Bay, Wisconsin | Refracting Reflecting | 102 cm 100 cm |
| Zelenchuskaya Astrophysical O. | Caucasus Mts., USSR | Reflecting | 600 cm |

* Six mirrors used together with the light-gathering equivalent of a single mirror this size.



U.S. Forest Service Photo by Roger M. Williams

Some astronomers believe that Wyoming's Big Horn Medicine Wheel could have been a prehistoric observatory. The unusual arrangement of stones could have served as sighting points for observations of the sun.

OBSERVATORIES

by John B. Irwin

An astronomical observatory is a station—a building or group of buildings—for the study of the sky. It houses the telescopes and other devices with which the astronomer works. Often we call a station of this type simply an “observatory,” but when we do so, the word “astronomical” is always understood. There are other kinds of observatories. Meteorological observatories are concerned with the weather; magnetic observatories, with the earth’s magnetism; seismological observatories, with earthquakes. These observatories are always known by their full names.

This article deals with observatories

that house optical telescopes. Some of these, plus other astronomical observatories, also include radio telescopes, antennae, and other equipment for the reception and analysis of radio waves and other invisible radiation from outer space. The study of this invisible astronomy is discussed in part in the article “Radio Astronomy.”

THE LOCATION

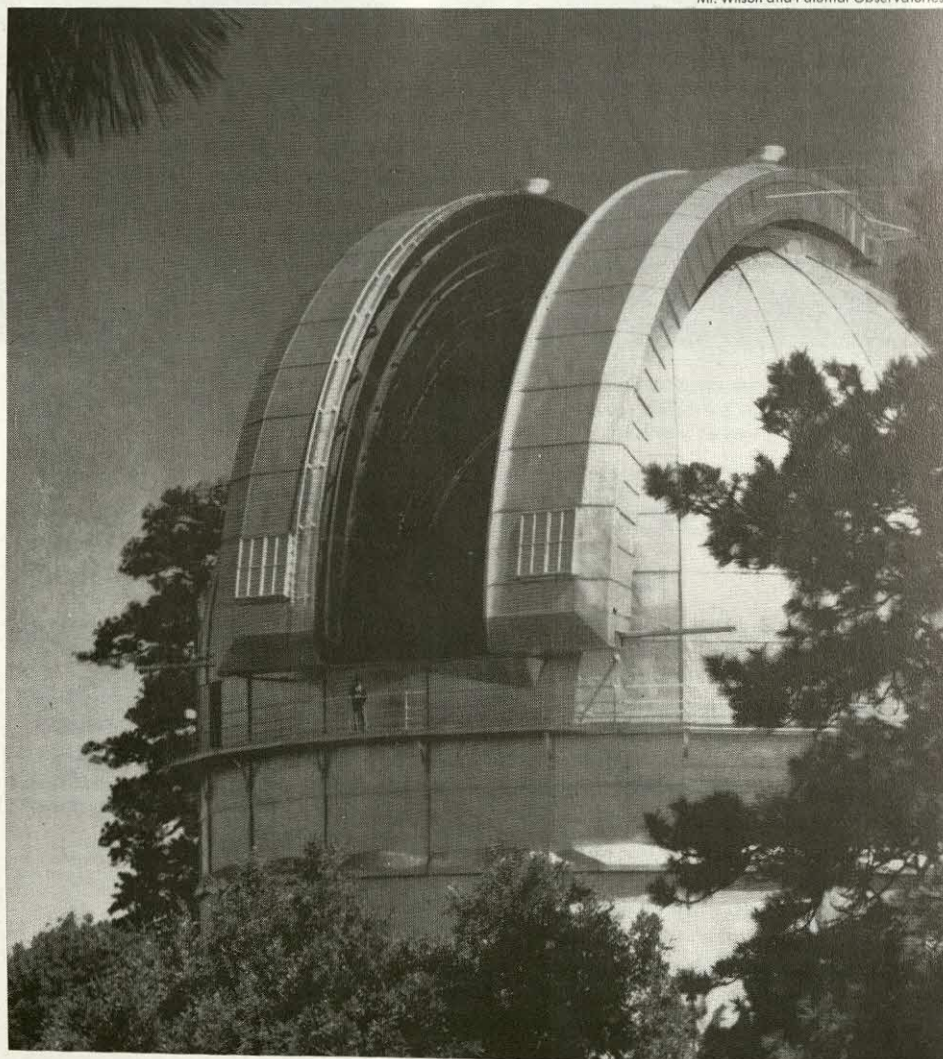
The site of an observatory devoted to optical astronomy must be as carefully selected as the instruments that will be housed in it. If the observatory is in a low-lying or comparatively low-lying area, the

image that appears in the lens or the mirror of a telescope will be faint and distorted. The reason is that the lower layers of air are often filled with ground fog or low-lying stratus clouds, as well as with a haze of dust or smoke. Nearness to a city is as much an obstacle to good viewing as low altitude. The lights of a great city can effectively blot out the fainter stars. Many an observatory originally built "out in the country" has had its effectiveness seriously reduced by the encroaching suburbs of a city.

For the best viewing, an observatory should be constructed on a height and in as isolated a place as possible. The English scientist and mathematician Isaac Newton pointed out the advantages of a high-altitude site. He wrote: "The Air through which we look upon the Stars is in perpetual Tremor . . . The only Remedy is a most serene and quiet Air, such as may perhaps be found on the Tops of the highest Mountains above the grosser Clouds." But even a mountain peak does not always make a per-

Mount Wilson Observatory in California. This photo shows a typical observatory with slit in the movable dome open to allow viewing through the telescope.

Mt. Wilson and Palomar Observatories



fect site for an observatory. Its crest, as any experienced mountaineer will tell you, is sometimes anything but "serene and quiet" and may produce its own storms. Yet when conditions are right, mountain observatories offer unparalleled viewing. Visitors to such observatories often marvel at the brilliance and great numbers of visible stars and are awed by the magnificent appearance of the Milky Way.

THE DOME

The first thing that strikes the eye as one approaches a typical large observatory is the white, hemispherical structure called the dome. There may be more than one. The telescope is housed within this structure. The longer the telescope, the larger the dome. The dome protects the telescope from the sun, rain, wind, dust, and heat. An opening, or slit, in it is kept covered when the telescope is not in use. It is opened up by means of sliding or folding panels when the astronomer is ready for observing. At night, as the telescope follows a star in its apparent rising and setting, the dome is rotated. It is driven by electric motors and rolls on a smooth track.

The dome is painted white with a special paint, so that it will reflect as much of the sun's heat as possible and not absorb it. The temperature must be kept the same inside and outside the dome. If it is hotter inside, the air will well up and "bubble" through the slit, thus spoiling the image that strikes the telescope lens or mirror. Sometimes large fans are located on the inner walls to circulate the cool night air inside the dome. Often a canvas or metallic wind screen is provided. This can be raised or lowered in the slit opening to protect the long barrel of the telescope from buffeting by the wind.

TELESCOPES AND THEIR MOUNTING

Telescopes are the focal points of observatories. Some of them look like gigantic spy glasses, with a large lens at one end of a long tube and an eyepiece (or, more commonly, a plate-holder) at the other. Such telescopes are called *refractors*. Their size is indicated by the diameter of the large

lens. The largest refractor in the world, at the Yerkes Observatory in Wisconsin, has a lens with a diameter of 102 centimeters, and a tube that is about 18 meters long.

Other telescopes—the *reflectors*—are quite different in shape. They have a great reflecting mirror set at the bottom of the large tube, which is not usually a closed cylinder but a skeletal structure. The starlight that is reflected from the mirror converges upward and comes to a focus—called the prime focus—at or near the top of the tube. As we shall see, the converging rays are made available for observation in several ways. The largest telescopes in the world are of the reflector type. The largest is the 600-centimeter telescope at the Special Astrophysical Observatory in the Soviet Caucasus mountains. The Hale telescope in California has a mirror diameter of 500 centimeters.

A large telescope, whether of the refractor or reflector type, is mounted on giant piers, solidly fixed in bedrock and built to a great height. The piers must be completely independent of the dome and the rest of the building. If the building were set to vibrating by the wind, by the motors, or by people walking about, the vibration would be transmitted to the telescope, and the image would be hopelessly blurred.

The telescope rotates about two axes—polar and declination. The polar axis is parallel to the axis of the earth. As our planet rotates from west to east, the apparent motion of the heavens in the opposite direction must be compensated for in order to keep a given celestial body exactly centered. The telescope is made to turn about the polar axis, by means of a motor or clockwork, at just the right speed to keep up with the stars. The declination axis is perpendicular to the polar axis. Rotation about the declination axis sets the telescope in the north-south direction.

ADJUSTMENTS FOR VIEWING

The astronomer looks through an eyepiece that may magnify 500 times or more. By means of small electric motors connected to gears, the astronomer carefully and frequently adjusts the telescope pointing.

The body that is being observed must be centered either on the magnified image of a pair of illuminated crosswires or on the narrow slit of a spectrograph, through which the light of the star is to be directed. A spectrograph is a device that spreads out, or disperses, light into its component wavelengths or colors, which are then recorded photographically. Each color corresponds to a definite wavelength.

In the domes that house large telescopes, a platform, or in some cases the entire floor, is raised or lowered by means of electric motors or hydraulic pumps to make the eyepiece comfortably accessible. In some cases, an adjustable observing ladder is provided.

In the case of certain reflecting telescopes, the starlight reflected from the mirror at the bottom of the tube is made to strike a small mirror at or near the top. From there, it is directed at a 90° angle to an eyepiece set at the side of the tube. This is called the Newtonian focus. The observer is stationed here on an observing platform that may be raised or lowered, as required.

The 500-centimeter Palomar reflector and the 300-centimeter reflector at the Lick Observatory near San Jose, California, are so large that it is possible for the observer to operate at the prime focus in a little cage centered in the top of the tube. The percentage of the starlight cut off by such a cage is not serious. As the telescope is focused on one star after another by making adjustments on a control panel far below, the astronomer in the cage may have a rather thrilling ride through the night.

There are more convenient systems for observing than those we have just described. In some telescopes, a small curved mirror is located near the top of the reflector tube. The mirror sends the reflected light from the mirror back again down the tube and through a hole in the main mirror. The light then reaches what is called the Cassegrain focus. This place—back of the big mirror and near the floor level—is much more convenient for observing than the Newtonian focus, near the top of the tube.

Even more convenient is the so-called coudé focus, at the lower end of the polar axis. It may take from two to four mirrors to get the light to this point, which is often in a temperature-controlled room. A very large, efficient, and stationary spectrograph can be located here.

The very large reflectors have a small field of view. They can sound the great depths of space in this direction and that, but they cannot give the complete overall picture. To supplement the work of such instruments, smaller wide-angle telescopes with an extended field of view are required. The smaller telescopes can do many jobs more quickly and effectively.

A telescope is a complex research device, which may serve in various ways. The earlier astronomers relied entirely on visual observation—that is, examining the desired object through the eyepiece of the telescope. This method is used only to a limited extent by modern astronomers. It has been largely replaced by various other types of observation.

TELESCOPE AS A CAMERA LENS

The telescope now very often serves as a giant camera, in which the telescope mirror or a mirror-lens combination—sometimes called a Schmidt telescope—is substituted for the camera lens. The light of a star or galaxy is directed to a photographic plate or film, where it is recorded. Much research astronomy is now conducted by this method. The entire sky has been photographed over and over again using different focal lengths, emulsions, and exposure times. A photographic plate can provide a permanent and easily duplicated record of a host of fine details in an object such as a galaxy—details much too faint to be seen by the observer stationed at the eyepiece.

Sometimes the light of a star or galaxy is not photographed directly, or even at all. A large telescope mirror may be used to gather up a large column of starlight and to focus the star—perhaps invisible at the eyepiece—onto a specially prepared alkaline surface. Through a photoelectric process, the light is converted into an electric current, which can be measured.



Richard Weymouth Brooks/PR

The Lick Observatory on Mt. Hamilton, California, one of the first observatories built in the United States, houses both reflector and refractor telescopes.

The colors of a star can also be precisely analyzed. A series of colored glass filters is interposed in the beam of light coming from the mirror of a reflector. Analysis of the colors helps determine the temperature and brightness of the stars. The reddening and absorption of light due to the dust between the stars and galaxies gives much valuable information about the composition of interstellar materials.

Every large reflector is used with spectrographs, which separate the incoming starlight into its component colors. For the brightest objects, the observable range of color may cover a large area of a photographic plate. For faint stars, the area is very small. By the analysis of spectrograms—photographic records made by spectrographs—the astronomer can obtain invaluable information about the composition of the stars, their motion in space, and other matters.

It must not be thought that the astronomer dispenses entirely with visual observation while doing direct photography, making photoelectric measurements, or using color filters or spectrographs. In all cases, the astronomer either looks at the object being studied to bring it into clear focus or else looks at a nearby guide or reference star.

Certain specially designed instruments are used at solar observatories. One of these is the *coronagraph*, which makes it

possible to observe the sun's corona, or outer layer, by producing the optical conditions of a solar eclipse. The telescopes used for studying the sun are much longer than ordinary telescopes and are usually housed in tunnels or towers. They cannot be moved. They capture sunlight through systems of mirrors that are trained to follow the sun. The solar telescopes at Sacramento Peak, New Mexico, and at Cambridge, England, are housed in vertical towers. On the other hand, the entire four-story building that houses the Multiple-Mirror Telescope at Mt. Hopkins, Arizona, rotates.

MANY TASKS

The astronomers working at observatories carry out a variety of tasks. At some observatories, they determine the correct time by observing the passage of certain stars across a given point. The Naval Observatory of the United States has such a time service. Every clear night, trained astronomers at Naval Observatory stations in Washington, D.C., and in Richmond, Florida, photograph the stars as they pass nearly overhead. After much measurement and calculation, the results of these observations are used to calculate "universal time," which is based on the rotation of the earth.

Astronomers also calculate exact latitude and longitude. International boundaries are fixed on paper by international treat-

ties but are actually marked out on the earth through the observation of the stars. To determine latitude and longitude requires stellar observations of the most delicate nature, together with the knowledge of the positions and motions of thousands of stars. The David Dunlap Observatory at Richmond Hill, Ontario, Canada, specializes in this work. It uses a 188-centimeter reflector telescope to calculate the positions and velocities of the visible stars.

The determination of time, latitude and longitude, and the calculation of star positions form what is known as fundamental astronomy. At the present time, this type of work is done in a comparatively small number of observatories.

Another important task of astronomers is to prepare photographic star maps showing the celestial bodies in the heavens. In 1887, under the auspices of the International Astronomical Union, about eighteen observatories all over the world began working on a map of the sky. Telescopic cameras of the same focal length and scale were used in obtaining the necessary pho-

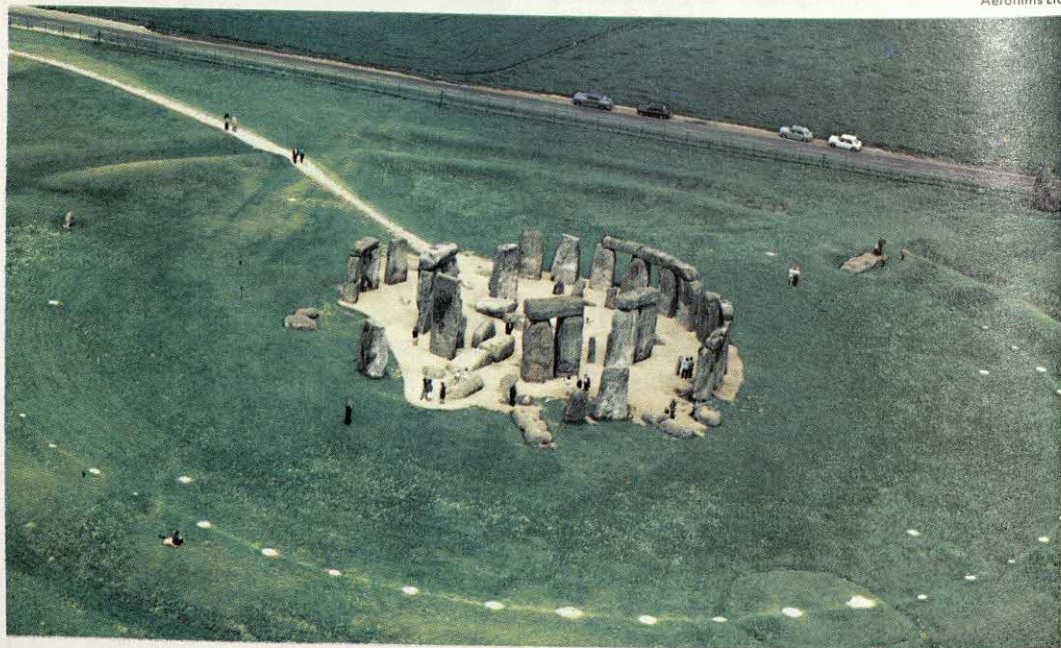
tographs. Work on this gigantic cooperative venture has now been completed. About 4,500,000 different stars have been recorded on more than 20,000 different photographic plates. The exact positions of a great many of these stars have been measured and recorded in catalogues printed in many languages.

Another photographic survey of the sky was done under the auspices of Mt. Palomar and the National Geographic Society. This survey was made with the 120-centimeter Schmidt telescope on Mount Palomar and recorded stars from the North Celestial Pole to -33° declination. Several hundred million stars to 21st magnitude have been photographed in this unexcelled survey. These magnificent charts are primarily used for identifications and statistical work.

Positional astronomers use photographs of zones of the sky taken with long focal-length telescopes to obtain accurate positions of the stars and other celestial objects. A series of photographs repeated several years later can provide the motions

Stonehenge may have been an early observatory. The stone arrangements can be used to calculate the position of the sun during an eclipse.

Aerofilms Ltd.



of celestial bodies. Perhaps the most important surveys of this type have been done by the Yale Observatory, in the United States, and by groups of cooperating observatories, in Europe.

Solar observatories study the surface features and the composition of the sun. They also conduct important research on the nature and quality of solar radiation. This radiation affects the ionosphere, the atmospheric layer that plays an important role in radio transmission. The Space Environment Laboratory of the U.S. National Oceanic and Atmospheric Administration issues monthly forecasts of ionospheric conditions based on the study of solar radiation.

There are many other kinds of pure research carried on at observatories. Among the fields of investigation are the proper motions, or apparent drifts, of the stars; their composition; their magnitudes, or brightnesses; the classification of their spectra; solar eclipses; the absorption of light by interstellar particles; and the general structure, chemical composition, and age of the universe. The Harvard Observatory published one of the largest catalogues of star brightnesses and star spectra.

In research carried on in an observatory, a great deal depends not only on how the stars are observed but on what stars are observed. For example, an astronomer with a 90-centimeter telescope equipped with a good photoelectric photometer can measure—with some effort—the magnitude and color of any of a hundred million stars. The astronomer must make a choice as to what few stars to observe. That choice is based partly on the astronomer's interests and training. It must also be based on what is already known and what other astronomers are doing. It may well be that halfway through an observing program, the astronomer may have to change it drastically because of some new fact that has turned up.

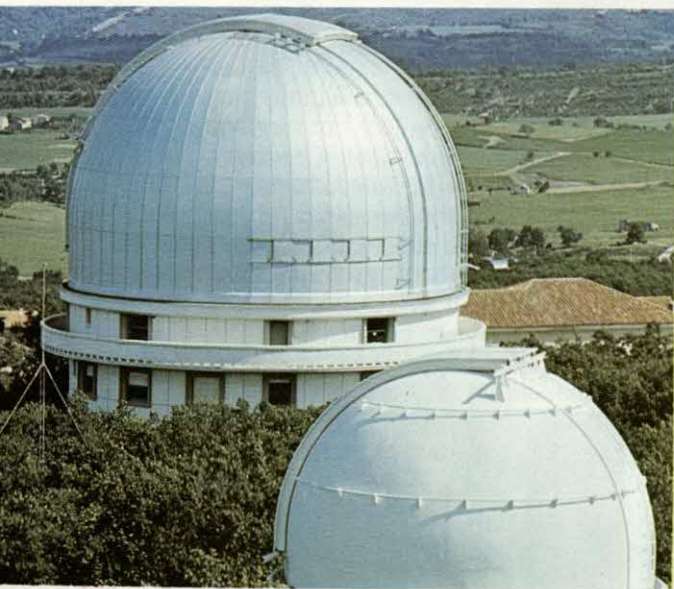
The 16th-century Danish astronomer Tycho Brahe in his observatory of Uraniborg. Brahe and his students catalogued the positions of more than one thousand stars.

THE DEVELOPMENT OF OBSERVATORIES

We have been dealing thus far with the modern astronomical observatory. We must bear in mind that it represents only the latest stage in a development that has been going on for thousands of years. Sites or posts for the observation of the heavens were used by the ancient Babylonians, Egyptians, and Greeks in the Old World and the Mayans in the New. The Greek astronomer Hipparchus, who lived in the

Josse-Grolier





A. Sole-Galliphot

Two European observatories: the Observatory of Saint-Michel-de-Provence in France, the largest observatory in Europe (above) and the Jungfrau Observatory in Switzerland (right).



U.P.I.

second century B.C., prepared a catalogue of the stars from observations he made from an observatory on the island of Rhodes. Generally, the observatories of the ancients were simply places where portable astronomical instruments could be used.

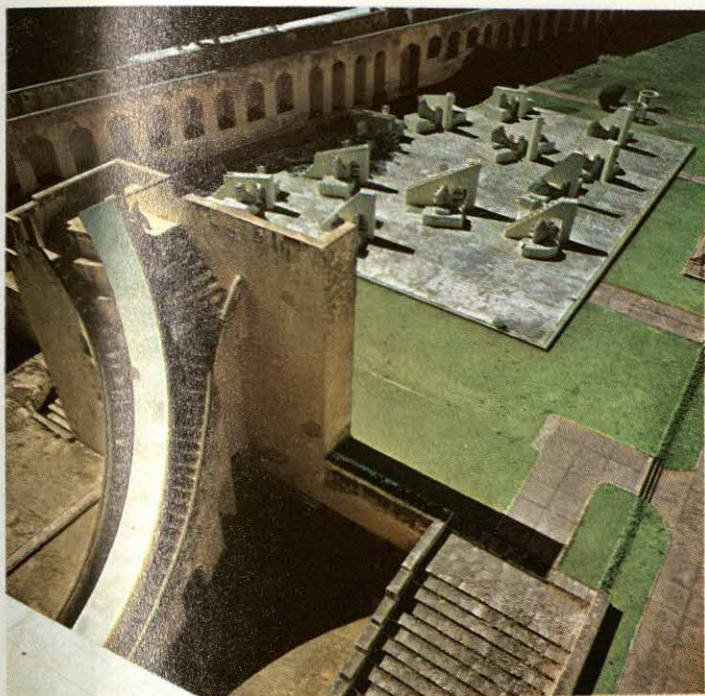
A renowned observatory was the one built for the 16-century Danish astronomer Tycho Brahe, on the island of Ven (or Hven), now in southern Sweden. The main observatory building was called Uraniborg (Castle of the Heavens). Tycho and his students did much important work here, including preparing a catalogue of the positions of more than one thousand stars.

Galileo Galilei was the first to use the telescope in astronomical observation. In 1609 he began to scan the heavens with the newly invented device. His "observatory" was the balcony of a building in the Italian city of Padua.

The first permanent observatory in the northern hemisphere was constructed at Greenwich, England in 1675. This observatory, established to provide a sound celestial coordinate system for navigators, is still in use today.

The Southern Hemisphere has fine observing conditions. North-central Chile was selected by ten American universities as the site for a 400-centimeter reflecting telescope, now in operation near La Serena at the Cerro Tololo Inter-American Observatory. The European Southern Observatory has a 366-centimeter reflector at Cerro La Silla, also near La Serena, Chile. The establishment of well-equipped observatories in the Southern Hemisphere has permitted the photographing of celestial objects not visible from Northern Hemisphere locations. The Cerro Tololo telescope was, for example, used to take the first modern photograph of 47 Tucanae, one of the brightest globular star clusters known.

The first permanent astronomical observatory in the United States was founded at the University of North Carolina, at



Michaud-Rapho

The observatory at Jaipur, in India, constructed at the beginning of the 18th century.

Chapel Hill, in 1831, though various temporary structures had been put up before that time. In 1875, James Lick, a San Francisco millionaire, left \$700,000 to the University of California for the establishment of a large observatory. It was erected on the top of Mt. Hamilton, in California. A 90-centimeter refractor was housed in the observatory. Later a 300-centimeter reflector was built there.

In 1917, a 254-centimeter reflecting telescope was erected in an observatory on Mount Wilson, California, by the American astronomer George Ellery Hale. An even larger telescope, with a mirror 500 centimeters in diameter, was installed in an observatory atop Palomar Mountain, California, in 1948. The largest of all reflecting telescopes, with a mirror almost 600 centimeters in diameter, has been built in the Soviet Union's Caucasus Mountains near the Black Sea.

A large observatory has been constructed on Kitt Peak, near Tucson, Arizona. Among its many instruments is the McMath solar telescope. Housed in a sloping, tunnel-like structure, it is the

world's largest solar telescope. It has a focal length of more than 90 meters.

The Kitt Peak National Observatory's management has an open-door policy that permits any qualified institution to set up its own observing facilities on the mountain. A new addition is the McGraw-Hill Observatory, run by the University of Michigan, Dartmouth College, and the Massachusetts Institute of Technology.

Canada has several outstanding observatories. The Dominion Astrophysical Observatory, established in 1916 at Victoria, British Columbia, has a 185-centimeter reflecting telescope. The David Dunlap Observatory at Richmond Hill, Ontario, has an even larger reflector, with a 188-centimeter mirror.

The earth's atmosphere blots out many wavelengths of electromagnetic radiation before they reach the ground. To study these wavelengths, which include X rays, ultraviolet light, and infrared rays, the United States has launched a number of astronomical observatories into space around the earth. Important optical observatories are listed in the table on page 44.

PLANETARIUMS

by Mark R. Chartrand III

Have you ever wondered what the sky looks like in another part of the world? Or perhaps what it looked like long ago, when ancient stargazers first started to study the patterns and motions of the stars? These puzzles can be solved and many other questions answered about the sky and the way it changes—all while you watch a fascinating show. Where? In a planetarium.

A modern planetarium allows you to see the stars, the planets, and the moon just as they appear in the sky and to watch as they change their positions. It also allows you to take a trip around the world, looking at the sky from different places. It can also give you a look into the past or a glimpse into the future. In New York City's Hayden Planetarium on a June day, for example, you can see not only what the sky will look like in New York that night, but also what it will look like in Rio de Janeiro. You can also see what it will look like in December in New York, or March in Rio. You can even see what it looked like the night you were born or what it will look like in 50 or 100 years.

Modern projection planetariums, such as the Hayden, are only a little more than 50 years old. However, the idea of making a model of the sky has been around for thousands of years.

EARLY MODELS

The earliest attempt we know of to provide a model of the sky was made in ancient Greece. The device was a sphere, surrounded by metal rings, each mounted so that it turned on a different axis. The sphere represented the celestial sphere, and the rings the paths of the sun and other heavenly bodies across the sky. Devices of this kind became known as armillary

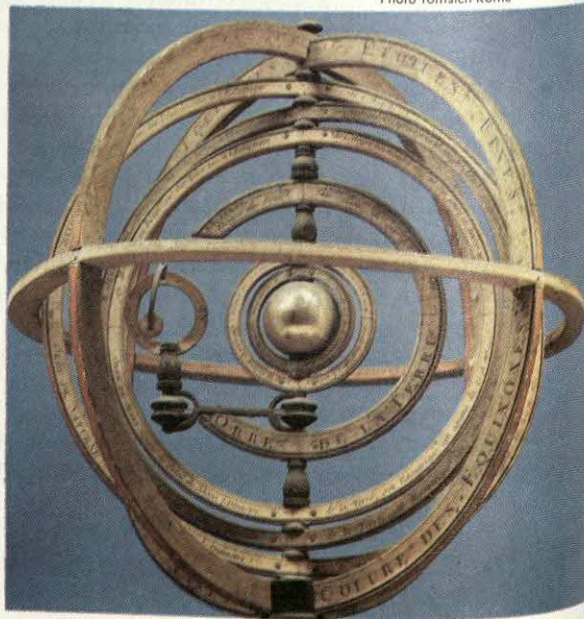
spheres. We do not know who made the first one, but the Greek astronomer Eratosthenes (about 250 B.C.) is often credited with the invention. We do know that Archimedes made a complicated model that was powered by water and accurate enough to reproduce eclipses of the sun and moon.

Another ancestor of the modern planetarium was the celestial globe. This was a

Scola



Photo Tomsich Rome



Early models of the sky. Top: an 18th century Copernican model showing the earth, moon, Venus, and Mercury revolving around the sun. Bottom: an 18th century armillary sphere.

sphere on which the positions of the stars and constellations were marked. Celestial globes were first made in ancient Greece. There was little further progress until about 1657, when a man named Andreas Busch made a large globe into which 12 people could climb. The stars were fixed on the inside of the globe, and there were rings for the planets to move along. More complicated globes were built in the years that followed. The last such globe was built early in the 20th century by Dr. Wallace Atwood for the Chicago Academy of Sciences, where it may still be seen. It is 4.5 meters in diameter and is turned by electricity.

Sometime after 1700, a different kind of mechanical model of the solar system was constructed. It reproduced the motions of the planets and showed the earth as just another planet. The model was made for the Earl of Orrery, and such devices are now called *orreries*.

A MODERN DESIGN

The modern planetarium did not develop until the 20th century. The idea for it came from Dr. Max Wolf of the Heidelberg observatory and Dr. Walter Bauersfeld of the Carl Zeiss Company of Jena, now in East Germany. They thought of using a very large hemisphere as a screen onto which the image of the sky could be projected. The hemispherical screen would be half of a celestial globe. The projection would be done by a special device that could imitate the appearance of the sky and

the motions of the sun, the moon, and the planets.

In August 1923 the first Zeiss planetarium, Model I, was shown in a makeshift dome on the roof of a factory. It was a single sphere with 31 projectors to show the stars and a number of smaller projectors to show the members of the solar system. It could only show the sky north of latitude 48° North, but the sphere could be turned to show the sky at any time of day or year.

Zeiss' Model II was a dumbbell-shaped projector, as all later Zeiss models have been. The latest Zeiss projector is the Model VI, which can be seen at the American Museum-Hayden Planetarium in New York City and at a number of other sites.

In 1947, Armand Spitz, an amateur astronomer from Philadelphia, designed and built a smaller, simpler, and less expensive planetarium. Spitz planetariums are now found in many schools and small museums that cannot afford the money or space for a larger instrument.

A MODERN PLANETARIUM BUILDING

The American Museum-Hayden Planetarium can be taken as an example of a major planetarium. The theater is 23 meters wide. Beginning 3 meters above the floor is a stainless steel dome whose highest point is 15 meters above the floor. In the center of the theater is the Zeiss projector. Arranged in circles around the projector are 750 seats.

The steel dome has millions of small

Gianni Tortolli/PR



Detail of the armillary sphere of Santucci of Pomerance. Sphere is made of wood and shows the planetary system and the northern hemisphere.

holes to allow the passage of air. They also make the dome lighter. In between this dome and the outer dome that can be seen from outside the building is a space that allows planetarium workers to place speakers and other devices for use in special shows.

THE PROJECTOR

A planetarium projector is designed to provide an accurate reproduction of the sky on the inside of the dome. It must be able to show all the stars visible to the naked eye. It must also be able to reproduce the motions of the planets around the sun, taking into account the changing point of view of the earth as it moves through space.

The stars are projected from the two spheres at the ends of the dumbbell. At the center of each sphere is a high-power light bulb. Arranged around the bulb are 16 lens systems, each containing a slide of an area of the sky. Tiny holes in the slides represent individual stars. Each hole lets light pass through, producing a point of light—a star—on the dome. The bigger the hole, the brighter the star's image. Holes to project the images of the brightest stars would have to be too large, however, so separate projectors are used for the brightest stars. Because these stars also usually show a slight amount of color, filters are used to

give the star image the correct color.

Projectors like those used for the brightest stars are also used to project images of the planets, and larger projectors are used for the sun and the moon. Some planetariums also contain mechanisms for showing eclipses of the sun and moon.

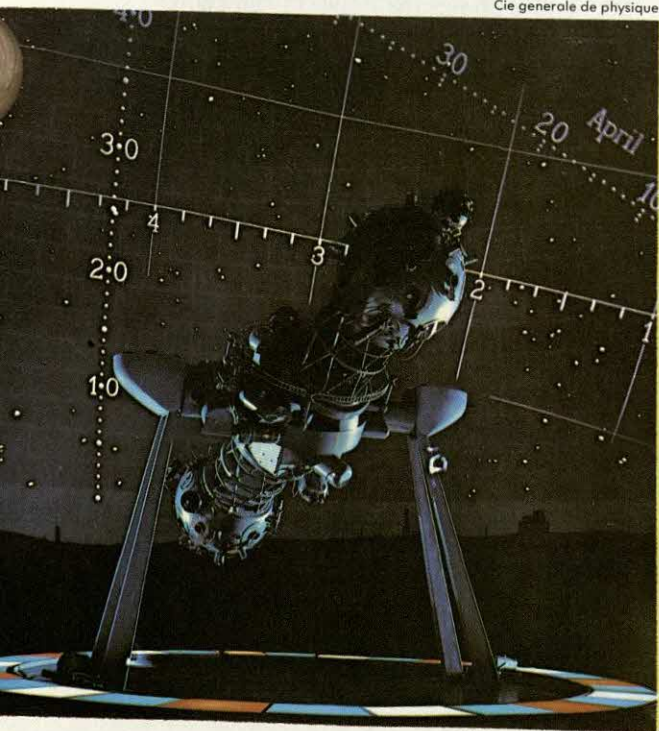
SHOWING THE SKY'S MOTIONS

A planetarium projector can show the sky as seen from any place on earth on any day in any year. It does this by means of complicated motors and gear systems that move the projections of the stars and solar system. These gear systems have been developed through detailed mathematical analysis of the paths of the heavenly bodies. Every planetarium can show four basic motions: daily motion, annual motion, precession, and latitude motion.

The daily motion is the sky's rotation around the celestial poles. This really represents the earth's daily rotation on its axis. Whereas the earth turns once in every 24 hours, the planetarium can go through this cycle in 30 seconds. Using daily motion, the planetarium operator can select the time of day to be shown on the dome.

The annual motion is the motion of the planets and the moon as they move around the sun. The projector operator can go through a year in one minute and can select

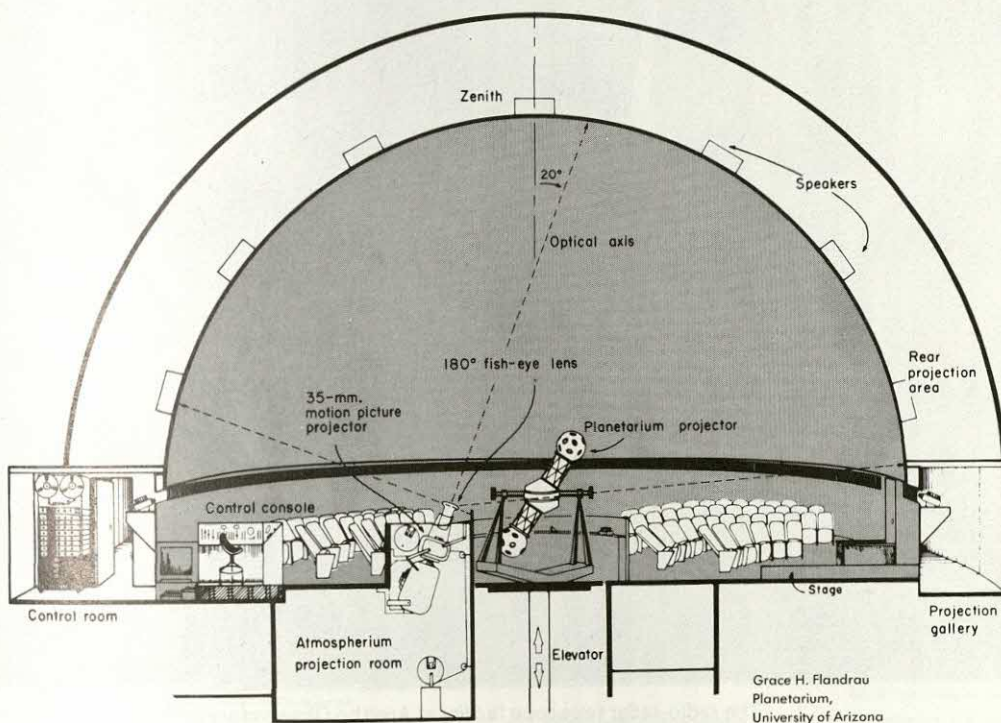
Cie generale de physique



The Zeiss projector. This device enables one to view the stars and planets as they appear at different times during the day or year, in either hemisphere. Below: close-up of optical system used in the dumbbell-shaped projector.

Fiske Planetarium, University of Colorado





A diagram of a typical planetarium, showing dome, seating arrangements, controls, and the Zeiss projector.

any day of any given year to show on the screen.

Precession is the slow wobbling of the earth's axis. It is similar to the wobbling of a top as it spins. Its effect is to change the position of the celestial north pole in the sky. The change in one year is very small, but over 26,000 years the celestial pole moves in a small circle in the sky. This motion is built into the gears so that as the planetarium is run using annual motion, the sky also precesses by the correct amount. An operator can go through the entire 26,000 year cycle in about one minute.

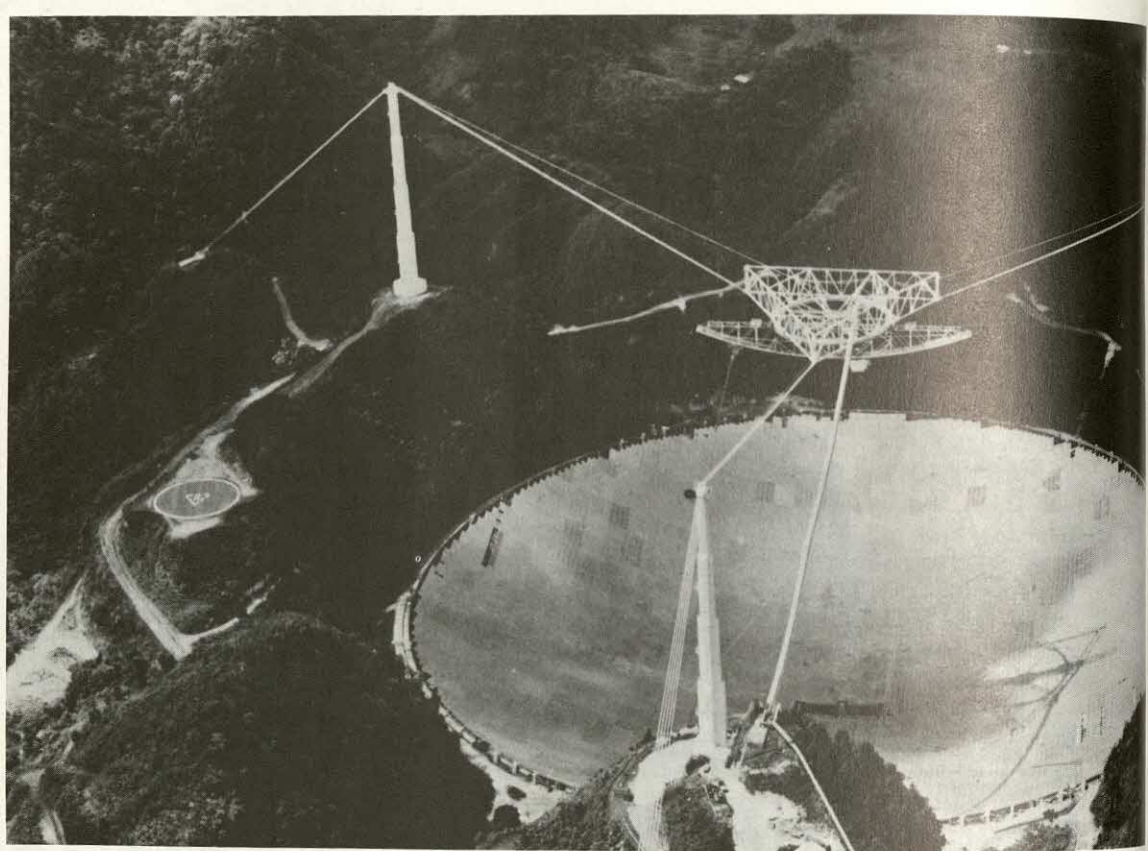
Latitude motion is rotation about an east-west horizontal axis. This allows the operator to set the part of the earth—New York City, Mexico City, or Rio de Janeiro, for example—from which the sky will be seen. One trip around our planet can be accomplished in one minute.

SPECIAL EFFECTS

Most planetariums also use special projectors to show such things as comets, eclipsing stars, supernovae, rocket launches, or other special effects. These projectors may be located under the main projector or elsewhere around the theater.

The operation of the entire planetarium is controlled from a panel near the side of the theater, sometimes with help from another person in a projection room just outside. The planetarium show is sometimes given live, and sometimes on tape with sound effects and music.

In these days of bright city lights and air pollution, most city dwellers get to see very little of the sky unless they travel away from urban areas. For these people, the modern planetarium is the best sky show in town.



The radio-radar telescope facility at Arecibo Observatory in Puerto Rico reaches farther into space than any other instrument.

RADIO ASTRONOMY

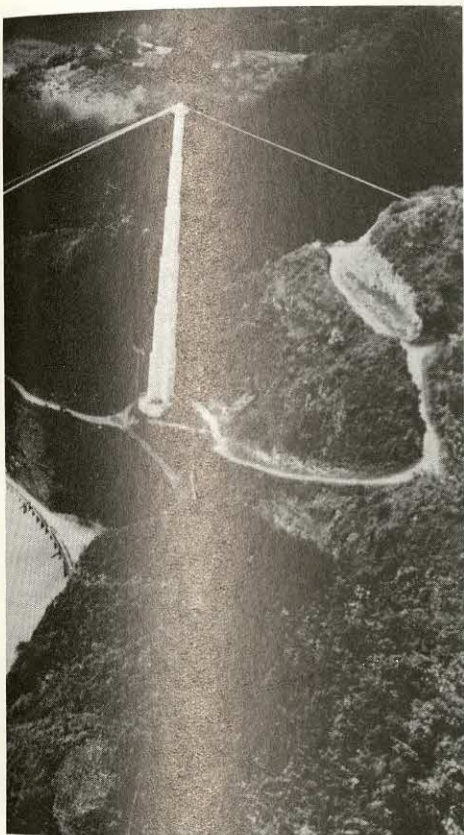
by Sir Edward V. Appleton

For centuries people have studied the heavens by way of the light waves from outer space which reach an earth-bound observer. At first, only the naked eye was used for the detection of such optical waves. Later, observations became more refined with the aid of optical telescopes and optical spectroscopes. Optical telescopes helped the observer to recognize faintly luminous objects and areas in the sky. Optical spectroscopes made it possible to break up any luminosity into its constituent colors.

In the mid-20th century, there developed an entirely new and parallel branch of astronomy in which the heavens are surveyed with a radio eye instead of an optical eye. This new science is called radio astronomy. It deals with the analysis of the

radio waves that are transmitted by certain heavenly objects and from certain areas in space. Radio telescopes and radio spectroscopes replace the optical telescopes and spectroscopes of the older visual astronomy.

Both visible light and radio waves are electromagnetic radiations, propagated through space at the speed of, roughly, 300,000 kilometers per second. Here one must note that the earth's atmosphere does not permit electromagnetic vibrations of all wavelengths to pass through it. An earth-bound observer is therefore denied the opportunity of studying the entire spectrum, or range, of the radiations which originate in outer space. Other forms of electromagnetic radiation, besides visible light and radio waves, included in this spectrum are



Cornell University Photograph

cosmic rays, gamma rays, X rays, ultraviolet rays, and infrared rays.

Fortunately, however, the absorbing atmosphere has two "windows," an optical window and a radio "window." It is only since the late 1940's that the radio window has been used for astronomical observations.

The band of radio wavelengths which passes through the earth's atmosphere ranges from about 0.25 centimeters on the short-wave side to about 30 meters on the long-wave side. This is, then, the wavelength band used by the radio astronomer. Mostly the astronomer examines the radiations (within this band) that actually originate in outer space. This examination has already revealed a great deal about the structure of the radio universe. Sometimes the astronomer generates radio waves and projects them towards celestial objects in order to detect them when they bounce back to the earth. We could appropriately call such reflection techniques *radar astronomy*. Both the moon and Venus have already been contacted by radar astronomy.



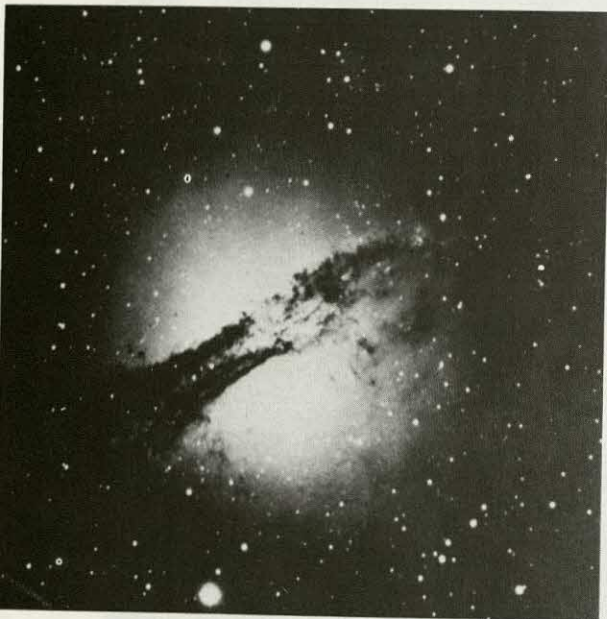
Cornell University Photograph

Footpads distribute a worker's weight as he walks on the paper-thin aluminum panels that cover the radio telescope's surface.

EARLY STUDIES

History tells us that both the American inventor Thomas A. Edison, in 1890, and the British physicist Sir Oliver Joseph Lodge, in 1894, attempted to detect radio emanations from the sun. We can now understand that their experiments failed because their detecting equipment was too insensitive. But great credit should be assigned to these pioneers for having conceived their experiments.

The most significant event in the history of radio astronomy was the discovery, in December 1931, of radio waves coming from the Milky Way by Karl G. Jansky, a radio engineer at the Bell Telephone Laboratories. Jansky was really studying static, that bugbear of radio. He found that two groups of static disturbances were due, respectively, to near and distant electric storms in the earth's atmosphere. But there remained a persistent hiss, the source of



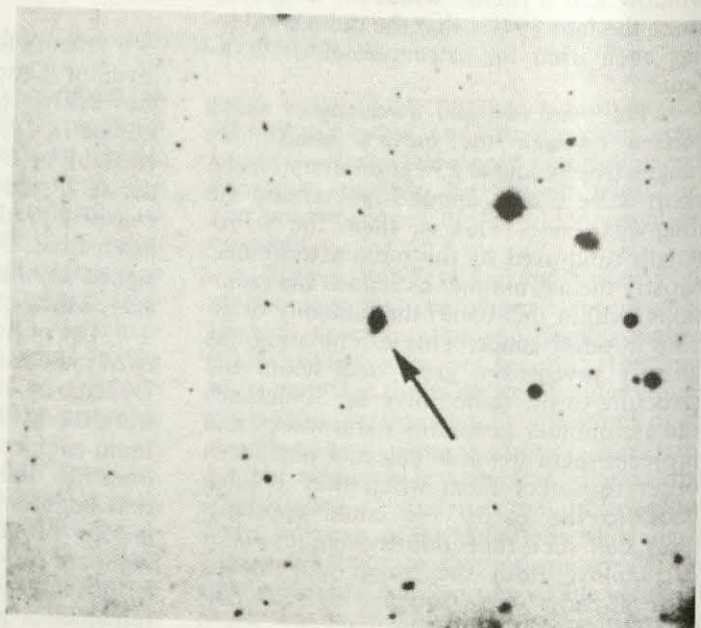
Mount Wilson and Palomar Observatories

Photo of a possible collision of two galaxies. Radio and optical astronomy complement each other. After radio emissions have been detected from a particular part of the sky, the optical astronomer knows where to focus his telescope and is then often able to discern the object and discover more about it. The radio source Cygnus A was optically found to be a double nebula, probably two galaxies in collision or interacting in some way.

which moved across the sky from east to west. At first the sun was suspected as the origin of this cosmic radio noise, but more careful tests showed that the waves came, instead, from the direction of our own galaxy, the Milky Way. Jansky's great discovery came about because he was able to recognize the unexpected, on which he then concentrated his attention until its true significance was revealed.

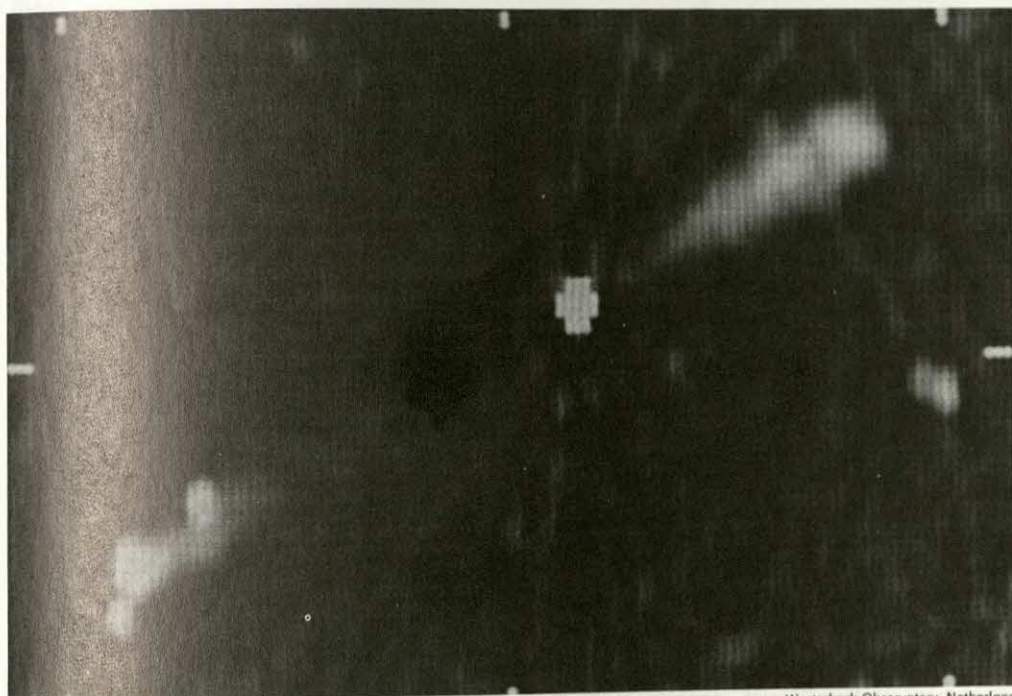
It will always be a mystery why Jansky's pioneer work did not at once prompt other investigators to seek to confirm and extend it. It is true that some valuable experiments with radio waves from outer space were made in 1936 by Grote Reber, a distinguished American radio amateur. He constructed a radio telescope, a device in which a large dish-shaped antenna brought to a focus the radio waves from a given area in the sky. With this device, which operated on a wave length of 60 centimeters, he mapped the skies. However, apart from Reber's contributions, radio astronomy lay quite dormant for a decade.

During the 1940's, cosmic radio noise was rediscovered, so to speak, but scientific work on it was postponed. Fortunately, much of the radio and radar equipment



Mount Wilson and Palomar Observatories

Object 3C-48 was the first radio source, other than the sun, that was identified as a single star. Other radio emitters proved to be diffuse nebulae or galaxies.



Westerbork Observatory, Netherlands

Largest known "object" in the universe shown in a photograph produced from its radio emissions. This object is 18.6 million light years wide.

which had been developed for military purposes during World War II turned out to be easily adaptable for some of the first experiments in radio astronomy. As the subject has developed, entirely new types of radio telescopes, often of large dimensions, have been designed. These new sophisticated telescopes, together with increased use of computer analysis, now permit greater accuracy in locating cosmic radio sources.

METEOR STUDIES

One of the first problems tackled in radio astronomy did not arise as a consequence of Jansky's discovery of cosmic radio. It is a matter of general observation that meteors, or "shooting stars," can be seen on cloudless nights. It has long been known, too, that these luminous streaks seen in the dark sky are due to pellets (generally tiny) of stone or metal which enter our atmosphere from outer space and burn up as they travel through it. Radar reflections can be obtained with the radio

telescope from the luminous streak of a meteor.

Two important results have been obtained from the examination of such radar reflections by the famous Jodrell Bank Radio-Astronomical Observatory, near Manchester, England. In the first place, some daylight meteor showers were discovered by these radio methods of detection. Such showers could not possibly have been discovered by optical means. Secondly, a radar method was developed which measured the rate of growth of the meteor trail and therefore the speed of the meteor itself. Scientists have long wanted to know whether meteors, the debris of space, belong to our own solar system or not. The Jodrell Bank radar measurements of meteor velocities all disclose values less than 72 kilometers per second—values lower than that required to escape the gravitational attractions of the solar system. It therefore looks as if we can picture all meteors as being members of the solar system, like the earth.



National Radio Astronomy Observatory

The largest radio telescope in the world, the Very Large Array, in New Mexico, consists of 26 antennas spread out over several square kilometers. Two of the antennas are shown above. The antennas are moved on tracks when necessary.

In subsequent years, extensive research in radio astronomy was undertaken. J. S. Hey, in England, used converted military radar equipment to extend the work of Jansky and Reber in mapping out the radio sky. Again he found the Milky Way to be a prominent and extensive radio source. In addition, he identified the first radio star—that is, a more or less sharply defined area of radiation in the sky.

RADIO STARS

The discovery of the first radio star, now called Cygnus A, was a momentous one indeed, for it was the starting point for much subsequent work. Radio astronomers have now identified some thousands of radio stars in the universe, but Cygnus A still appears to be the most powerful of them.

The Milky Way contains a number of radio stars. Practically all the radio stars discovered so far have not been individual

IMPORTANT RADIO TELESCOPES

Paraboloids (singles)

| Location | Diameter (m) |
|-----------------------|--------------|
| Arecibo, Puerto Rico | 305 |
| Bonn, W. Germany | 100 |
| Green Bank, W. Va. | 92 |
| Haystack Hill, Mass. | 37 |
| Jodrell Bank, England | 76 |
| Nancay, France | 305 |
| Parkes, Australia | 64 |
| Pulkovo, USSR | 150 |
| Stanford, Calif. | 46 |

Crosses (two or more telescopes)

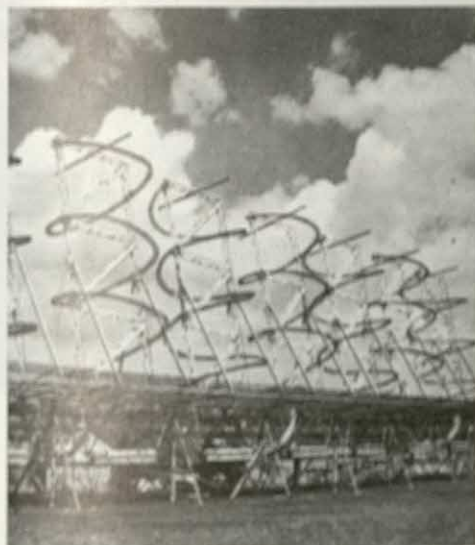
| Location | Length (m) |
|------------------------|------------|
| Bologna, Italy | 1,000 |
| Cambridge, England | 420 |
| Fleurs, Australia | 550 |
| Hoskinstown, Australia | 1,600 |
| Serpukhov, USSR | 1,000 |
| Stanford, Calif. | 110 |

stars, but rather galaxies in collision, luminous gas clouds, or the remnants of exploded stars. In January 1961, the California Institute of Technology reported that it had located the first radio star, other than the sun, that was really a single star. This object, called 3C-48, is in the constellation Triangulum. Dr. Jesse L. Greenstein of the California Institute of Technology noted that this discovery "marks the first time astronomers have been able to listen to and locate a star that emits a radio signal."

Let us consider for a moment our own radio star, the sun itself. Because it is a hot body, the sun emits thermal (heat) radiation in the radio spectrum. The strength of such radiation is determined by the solar temperature. But, in addition, the sun emits non-thermal radiation from time to time, often in the nature of bursts of energy. The largest solar radio storms of this kind are often associated with sunspots and solar flares, as



University of Ohio



University of Ohio

was first shown by E. V. Appleton and J. S. Hey. Elegant experiments carried out in Australia later showed that the actual source of a radio-noise outburst is a stream of corpuscles, ejected from the solar atmosphere at the time of a solar flare. This stream produces auroras and magnetic storms later when it reaches the earth's atmosphere. When the sun's quiet thermal radiation is measured, its intensity is such as to indicate a temperature of 1,000,000° Celsius. Yet the disk of the visible sun is known to have a temperature of only 6,000° Celsius. The explanation of the discrepancy is that the thermal radiation detected by the radio astronomer comes from the solar corona, or outer layer, which is much hotter than the solar disk.

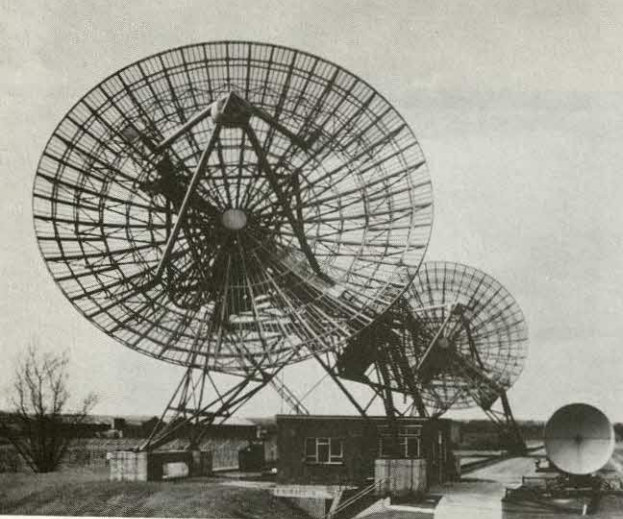
What the Milky Way would look like if our eyes were sensitive to radio waves of 1.2 meters. The very bright area (lower right) is the central region, near the nucleus, of our galaxy. The smaller bright area above is the region of Cygnus and the beginning of one of the spiral arms of the galaxy. The sun is embedded in this spiral arm. This view is based on a radio map of the galaxy prepared by radio astronomers at Ohio State University's radio telescope facility (left).

Using their radio telescopes, radio astronomers have detected thermal radiations in the radio spectrum from the moon, Venus, Mars, Jupiter, and Saturn, which enable us to tell how hot these bodies are. Only in the case of Jupiter does it appear that the constant thermal radiation is sometimes accompanied by short-period high-energy bursts, which can be detected in the wave-length range of 15 to 20 meters. The origin of these violent radio storms is still uncertain.

The most basic contributions to human knowledge yet made by the radio astronomer have been the identification of the positions of thousands of radio stars. This work began, as we have seen, with the identification of Cygnus A by J. S. Hey. Soon thereafter, the second strongest radio star, Cassiopeia A, was discovered by M. Ryle and F. G. Smith. Ryle and his colleagues at Cambridge University have been prominent in conducting radio-star surveys. Other workers in the field have been B. Y. Mills of Australia and J. G. Bolton, formerly of Australia and now of California.

WORKING WITH OPTICAL ASTRONOMY

The universe disclosed in a radio-star survey does not tally in detail with the more



John T. Scotts, *Physics Today*

Portion of the Cambridge University radio telescope used by Nobel laureate Sir Martin Ryle in some of his early work in radio astronomy.

familiar universe of optical astronomy. Often the strongest radio emissions come from objects in the sky which are only faintly discernible by optical methods. However, radio telescopes are limited in their ability to distinguish between two objects close together. The famous Cygnus A source was only identified with a visible object after the optical astronomer was told precisely where to look with his powerful telescope. It then turned out that the powerful radio emission came from a faint double-structure nebula, which is now known to be two galaxies in collision. The light and radio waves emitted by Cygnus A take 550 million years to reach the earth.

Another, and yet entirely different, type of radio source which has been identified optically is the Crab Nebula. As far back as the year 1054 A.D., Chinese observers witnessed and recorded the remarkable occurrence of the explosion of a star, which brought about a greatly increased luminosity. Such an exploded star is called a supernova. The Crab Nebula is the best-known example of a radio transmitter.

Nowadays it is suspected that a number of other radio stars besides the Crab Nebula are remnants of supernovae. This is certainly the case with the radio source Cassiopeia A, which is a member of our own galaxy. Tycho Brahe, in 1572, observed a "new star" at the position of this

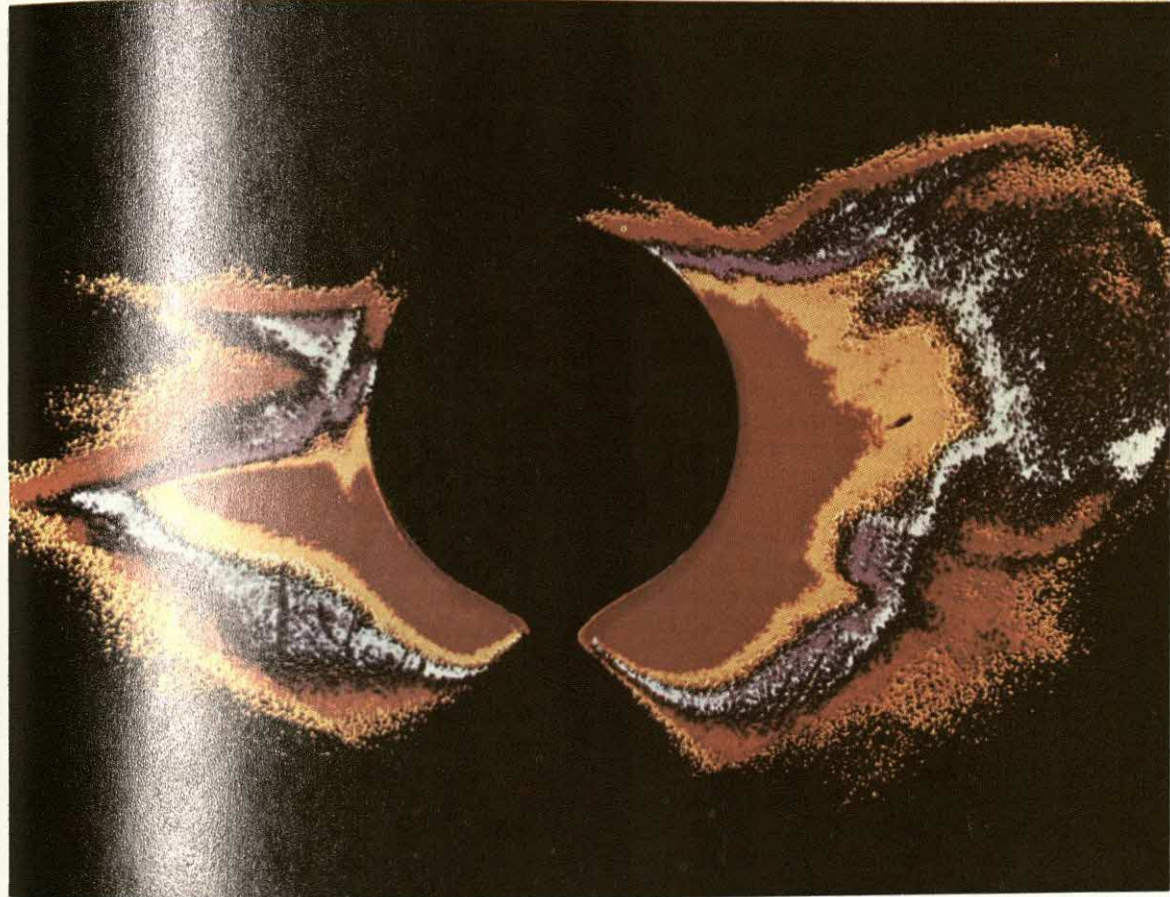
radio source. Therefore, what he observed might have been a continuing explosion that was changing a star into a supernova that emits radio waves.

RADIO COSMOLOGY

The attempt to associate radio stars with visible objects is still continuing, sometimes with success and sometimes without. The radio astronomer has felt encouraged to study the radio universe as it reveals itself, without reference to the visible universe. In this way a new subject has developed: that of *radio cosmology*.

A central problem of radio cosmology is that of determining the distribution of radio stars out to the farthest limits of space. The distribution of such stars has an important bearing on the two rival theories of the origin of the universe. One theory holds that matter is being continuously created and that the distribution of stars should be uniform everywhere. According to the rival theory, the universe has evolved from a tightly packed nucleus which exploded outwards ten thousand million years ago. It is maintained that the stars are not uniformly distributed but show a greater concentration at greater distances. The bearing of radio-star surveys on this basic question of the uniformity or nonuniformity of the radio universe is being intensively examined at present. No one would claim that a final pronouncement can yet be made. But the evidence appears to favor the view of the Cambridge radio astronomers that the more distant radio stars are more closely packed than the nearer ones. This result would seem to support the evolutionary theory of the universe.

Mention must be made of the radio waves emitted by hydrogen atoms in space. That such atoms would emit with a wave length of a little over 21 centimeters was first predicted by H. Van der Hulst, in Holland, some years before the phenomenon was actually detected. Studies of such radiation have helped radio astronomers to map out the distribution of hydrogen in our galaxy and to analyze its movement in space. All this supports the view that the Milky Way has a spiral structure.



NASA

With the sun's disk artificially masked, we are able to see the sun's outer atmosphere, or corona, extending outward for millions of kilometers. The photo has been color coded, each color distinguishing different levels of brightness.

THE SUN

by Oran R. White

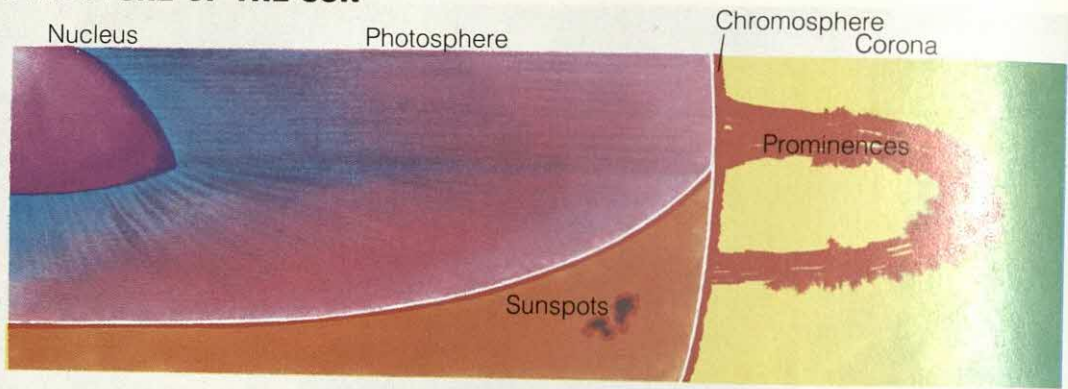
The sun is the center around which the earth and the other planets of the solar system ("solar" means "relating to the sun") revolve. Many ancient peoples around the world worshiped the sun. They made offerings and built temples to the sun. For the most part, their life depended upon agriculture, and they associated the sun's warmth with the growing season.

Although the ancient sun worshipers knew much less than we do now about the sun, they sensed its life-giving importance. Our sun is indeed crucially important to us. It provides the earth with the heat and light necessary to sustain all living things.

The sun is a star—a rather ordinary one, in fact. Our sun is only of average size. Many other stars are bigger, heavier, hotter, and brighter. The sun appears to be much bigger and brighter because it is much closer to us than any other star is. It is about 149,600,000 kilometers away. The next nearest star, Alpha Centauri, is more than 40,000,000,000,000 kilometers away.

Modern astronomers have learned that our sun is only one of about 100,000,000,000 stars in our star group, or galaxy, called the Milky Way. The sun and its family of planets are located in one of the spiral arms of the Milky Way, at a

STRUCTURE OF THE SUN



point about three-fourths the distance from the center to the edge of this galaxy.

PHYSICAL DATA

We cannot scorn the reverence with which earlier peoples regarded the sun. Modern knowledge confirms that the sun is awe-inspiring in measurable ways—in its size, its mass or weight, its densities, its pressures, its temperatures, and many other features.

Size. The sun is a vast ball of unbelievably hot, glowing gas. It is some 1,400,000 kilometers across—more than 100 times the diameter of the earth.

We would expect a globe the size of the sun to be stupendously heavy. The sun's mass, in fact, equals that of 333,420 earths. Because of the weight of this vast amount of gas, the pressure at the center of the sun is more than one million metric tons per square centimeter.

Density. In spite of the great mass of the sun, its average density—the weight of a standard volume of its matter—is only 1.4 times the weight of an equal volume of water. The earth, on the other hand, is 5.5 times denser than water.

This low solar density is easy to explain. The center of the sun, because of the enormous pressure, is more than 100 times denser than water. But much of the sun beyond the center is composed of gas that is often thinner than the earth's atmosphere. When these densities are averaged together, the general density of the sun is quite low.

Gravity. Because of its great mass, the sun has a gravitational pull 28 times stronger than the pull of the earth. This means that if a man weighing 90 kilograms on the surface of the earth were put down on the surface of the sun, he would weigh 28 ×

90 kilograms. This equals 2,520 kilograms, or about 2½ metric tons. The person would not be concerned about his weight increase. He would vaporize instantly, because of the heat.

Temperature. The sun is like a huge furnace, fired by nuclear, or atomic, energy at its core. Temperatures at the center may be 14,000,000° Celsius or more. At the surface of the sun, however, temperatures are much cooler—between 5,000° and 6,000° Celsius. This is still hot enough to vaporize nearly all substances that exist as solids or liquids on the earth.

Structure. The sun is composed of several distinct regions. It has an atmosphere, consisting mainly of two layers. Below the atmosphere is the surface, which is called the *photosphere*. *Sunspots* are an important feature of the surface. The interior and core of the sun are almost unimaginably hot, as we have seen. Although scientists cannot easily investigate the sun's interior directly, they have developed theories about it from what they do know. We shall now examine these structural features of the sun.

ATMOSPHERE OF THE SUN

Extending far upward from the surface of the sun is the solar atmosphere. The atmosphere consists mostly of hydrogen gas. It is much less dense than the rest of the sun.

The solar atmosphere is composed of two layers. The lower, or inside, layer is the *chromosphere*, or "sphere of color". It extends as much as 12,000 kilometers above the sun's surface. The higher, or outside, layer is the *corona*, or "crown". The corona forms a beautiful white halo around the entire sun, sending long streamers millions of kilometers out into space.

We on earth cannot usually see the chromosphere and the corona. The effects of our own atmosphere and the bright glare of the photosphere blots them out. But the solar atmosphere becomes visible during a total eclipse of the sun, when the moon covers the photosphere.

We do not have to await an eclipse to observe the corona, however. We can see it through a special telescope called a *coronagraph*, which produces an artificial solar eclipse. Or we could travel to outer space, above the earth's troublesome blanket of air, where the corona is always visible to the unaided eye.

Temperatures in the corona and the chromosphere vary in an unexpected way. The lower part of the chromosphere may be less than 5,000° Celsius, which is cooler than the photosphere. But the temperature rises in the outer reaches of the chromosphere, reaching 10,000° Celsius or perhaps even 100,000° Celsius in the uppermost levels.

The corona is much hotter than the chromosphere. Astronomers have estimated an astounding temperature of 2,000,000° Celsius for its outer reaches. Why the corona, which is so far from the source of the sun's energy, should be so much hotter than the photosphere is rather puzzling. One explanatory theory holds that strong shock waves, caused by turbulent movements of the photosphere, heat the very thin gases of the corona intensely.

A spectacular activity occurring in the chromosphere is known as the *prominences*. Prominences are huge streamers of glowing gas, sometimes reaching heights of hundreds of thousands of kilometers into the overlying corona. They take a great variety of shapes. Prominences are best observed during a solar eclipse or with a coronagraph.

Some prominences are eruptions or explosions, rising quickly and soon fading away. Other types last much longer. Still other prominences seem to originate high in the chromosphere and then rain gas downward, toward the sun. The occurrence and lifetimes of prominences are influenced by solar magnetic fields nearby.

The chromosphere also displays much smaller jets or filaments of gas, called *spicules*. These may result from strong movements of hot chromospheric gas. These gas movements appear in the chromosphere as coarse cells, called *supergranulation*.

From time to time, the chromosphere is active in other ways. Hot, bright markings—*plages* and *flares*—are often observed. Plages are areas of hot brightness. Flares are high-energy outbursts of radiation and tiny, subatomic particles. These particles may reach the earth's atmosphere.

SURFACE OF THE SUN

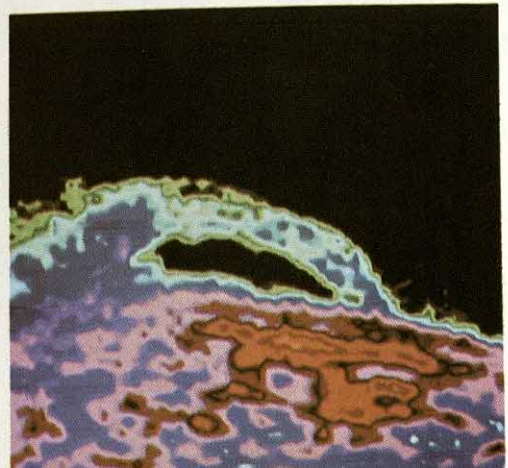
It may seem strange to talk of the surface of a gaseous globe like the sun. But there is a surface layer, which has a definite border and is the deepest part of the sun that we can see. This layer is called the solar disk, or the photosphere ("globe of light").

The photosphere is a relatively thin region. It has a depth of some 320 kilometers. This is less than $\frac{1}{2,000}$ of the radius of the sun.

The photosphere was once thought to be a uniform and perfect orb of light. But even in ancient times observers occasionally saw spots on it. In the early 1600's the Italian scientist Galileo Galilei became the first man to study the sun and its spots through a telescope. These so-called sunspots are dark, irregular patches. They are thought to be very important but are in many ways still mysterious features.

The chromosphere is a site of spectacular activity. Here huge streamers of glowing gas reach into the corona in a form known as a limb prominence.

NASA and Harvard College Observatory



In addition to sunspots, the solar surface shows two other main features: bright irregular areas called *faculae* and a network of fine cells, the *photospheric granulation*.

Faculae, or "little torches," are hot, glowing regions, ranging from tiny bright marks to huge splotches. They resemble the plagues in the chromosphere. They often surround sunspot groups, but they may occur alone. Faculae often arise where sunspots later appear, and then last for a while after the sunspots have vanished. Faculae have a coarse-grained structure. Many astronomers consider them to be huge masses of gas that are hotter than the rest of the gaseous solar surface.

Through the telescope, photospheric granulation looks like bright grains of rice. The grains are separated from one another by dark boundaries. A typical grain, or cell, measures about 1,600 kilometers across—actually a small area compared with the enormous surface area of the sun.

Astronomers consider the granulation to be photospheric gas in continuous and violent motion because of heat. Movies have been taken of the cells, which look like boiling fluid bringing up gas from the depths of the sun.

SUNSPOTS

Sunspots appear dark because, at a temperature of 4,000° Celsius, they are cooler and thus less bright than the rest of the photosphere. A typical sunspot has two distinct parts: a dark central region, called the *umbra* ("shadow"), and a lighter surrounding area—the *penumbra* ("almost shadow"). Like the photosphere generally, the umbra shows a granular structure, which suggests the circulation of hot gas. At certain positions near the edge of the sun, the spots look like depressions, or hollows, in the photosphere.

Near right: a map of the sun's corona, presented on a computerized color display. The north and south poles are black; the hottest regions are white. The map was obtained by an Orbiting Solar Observatory (OSO). The illustration on the far right shows an OSO. A coronagraph, an instrument that creates an artificial eclipse for the sake of studying the sun's corona, protrudes from the front of the satellite.

A single spot may be tens of thousands of kilometers wide. It is usually a temporary feature, lasting anywhere from a few days to a few months. Sunspots often develop in pairs, which then tend to drift apart slowly. Over a period of time, hundreds of spots may form in large groups. At still other times, there may be practically no sunspots at all.

Many theories have been advanced to explain the nature of sunspots, but none of them tell us everything we want to know. Over the years, the spots have been compared to low-pressure areas or tornadoes or huge whirlwinds in the solar gas. And, in fact, complex movements of gas both into and away from a sunspot have been observed.

More recent theories hold that sunspots are cool areas produced by reactions between the electrically charged gases of the sun and solar magnetic fields. That is, a local magnetic field breaks through the surface of the photosphere, producing a spot at that point. This disturbance also affects the solar atmosphere overlying the spot.

NASA



Sunspots actually do have powerful magnetic fields. In a sunspot pair, one spot has the positive (+), or north, magnetic pole; the other, the negative (-), or south, pole of an associated field. Sunspot magnetism may be related to electrical currents passing through the solar gas.

SOLAR CYCLE

Sunspots appear and then disappear in a definite cycle, called the *sunspot cycle* or the *solar cycle*. The average duration of the cycle is 11 years. As the cycle begins, there are only a few small spots on the face of the sun. As the cycle progresses, spots become more numerous.

A sunspot cycle begins with a few small spots at solar latitudes of 30° to 40° north and south of the sun's equator. With the passage of time, more sunspots of larger size appear. These new spots arise closer and closer to the solar equator and to the poles, until much of the sun is covered by dark patches.

The cycle nears its end when spots in the higher solar latitudes begin to vanish.

At last only a few spots are left around the equator. Then the spots of the next cycle start to appear, at 30° to 40° north and south of the solar equator.

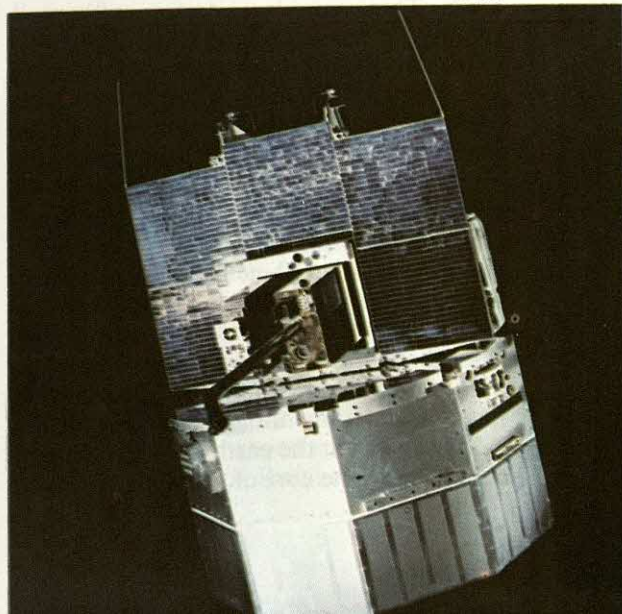
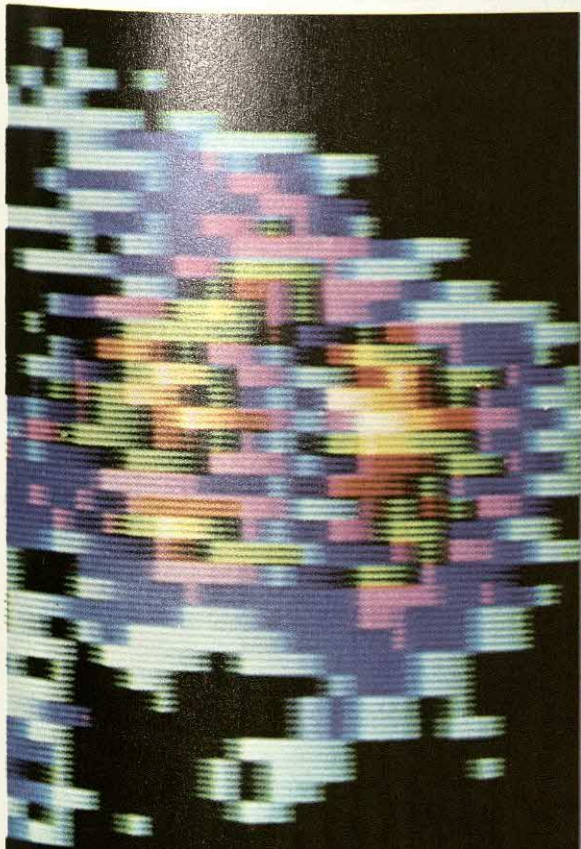
Some astronomers have linked the spot cycle to a complex circulation of solar gas: from the surface of the sun down and then up again, and from the solar poles to the equator and back. The heat and rotation of the sun are supposed to produce this effect. The circulation has been compared to that of the earth's winds and ocean currents.

More modern sunspot theories make use of new discoveries about magnetic fields. Scientists have found that a hot, electrically charged fluid, such as the solar gas, produces magnetism. As the gas moves, the lines of magnetic force follow. Regular movements of solar gas and its accompanying magnetic fields may cause the sunspot cycle.

Astronomers have discovered that the north and south magnetic poles in the sun switch positions in a regular, repeated fashion. These changes take place at the beginning of each sunspot cycle. The result is a magnetic cycle of 22 years, or double the duration of the sunspot cycle. This seems to indicate that the spot cycle is definitely connected with magnetic forces in the sun. But exactly how is a mystery.

Sunspots are not the only features affected by the solar cycle. During a *sun-*

NASA



spot maximum, when many spots exist, the entire sun becomes more active. Prominences are larger and more common. Huge solar flares, with temperatures in the millions of degrees, burst forth.

The corona undergoes changes in shape and brightness during different phases of the solar cycle. Man-made satellites have also photographed a strange feature, known as solar polar caps, in the corona. These coronal caps, centered over the north and south poles of the sun, are relatively cool masses of gas. At about 1,000,000° Celsius, the caps are only about half as hot as the rest of the corona. During a sunspot maximum, the polar caps are very small or absent. During a sunspot minimum, when only a few spots exist, the caps become much larger.

During sunspot maximums, other parts of the corona become about twice as hot—4,000,000° Celsius—as usual, especially where solar flares take place.

The activities associated with sunspot cycles often affect the earth and its atmosphere. These effects are due in large part to the many electrically charged, high-speed particles and, also, ultraviolet, and X rays emitted from the sun during a sunspot maximum. Many solar particles are trapped in the earth's magnetic field. Others reach the earth's atmosphere, where they cause the bright lights called *auroras*, around the north and south poles. The particles and rays may create magnetic storms on earth. They can also interfere with compasses, communications, and electrical power transmission.

INTERIOR OF THE SUN

There is little direct evidence available to scientists concerning the inside of the sun. Astronomers studying the sun know mostly about the heat, light, magnetism, and particles that come from the upper layers of the photosphere and solar atmosphere.

There are, however, some clues to the nature of the sun's interior. Scientists believe that certain atomic particles, called *neutrinos*, reach the earth directly and very quickly from the core of the sun. These par-

ticles are small. They have no charge and hardly any mass to speak of. They move at the speed of light. Because of these properties, neutrinos easily pass through great thicknesses of matter, making them hard to detect. Nevertheless, scientists have managed to detect and count some of the solar neutrinos.

From neutrino research and from what is known about the outer layers of the sun, scientists have constructed the following theoretical model of the solar interior.

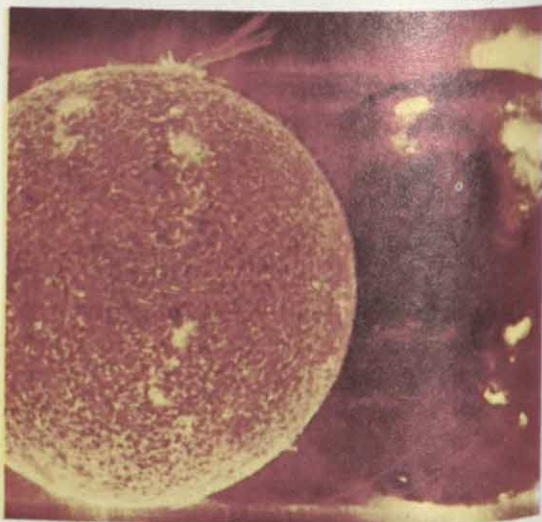
A "nuclear-energy furnace" probably forms the sun's core. It fills a relatively small volume. The core is extremely hot and very dense. The atoms there have lost their electrons and consist only of nuclei.

Circulating currents bring hydrogen—atomic fuel—to the "furnace" and carry away the resulting product—helium. The energy released must somehow escape from the core. Otherwise the sun would swell up and explode.

The heat, light, and other energy from the core are slowly radiated and reradiated from atom to atom. Gradually the energy is transferred outward, away from the core. As the energy reaches the sun's surface,

A solar eruption is visible at the top of the left image in this sequence, which was taken by the U.S. Skylab crew.

NASA



another form of energy transfer takes over. Circulating currents of solar gas carry the energy.

Like boiling water or hot air shimmering over a fire, the gas of the photosphere rises and falls from the energy of the heat. These photospheric movements produce the many phenomena seen at the sun's surface: granulation, faculae, and spots.

Scientists estimate that it takes several million years for energy other than neutrinos from the solar atomic core to reach the surface of the sun. From the solar surface, this energy radiates in all directions.

INSTRUMENTS FOR OBSERVATION

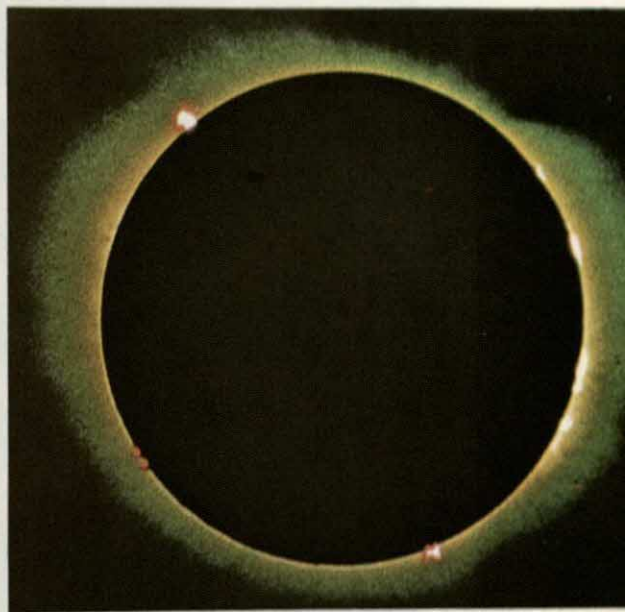
Astronomers find the sun interesting because as yet it is the only star they can examine relatively close at hand. It has given them important clues to what the nature of many other stars may be. But astronomers must use instruments in their study—not only to magnify the sun but also to protect their eyes. One should never look directly at the sun, either with the naked eye or through an ordinary telescope or binoculars, because the sun's rays may damage the eyes, and blindness could result.

Instruments for observing the sun have existed for many hundreds of years. Before the invention of the telescope and other modern astronomical devices, people used simple mechanical solar instruments to measure the positions and paths of the sun.

Most modern solar telescopes are very different from ordinary astronomical telescopes. A solar telescope may be very large and complicated. One kind of solar telescope is called a *tower telescope*, because the mirrors and lenses for concentrating sunlight are located atop a tower.

A tower telescope projects an enlarged image of the sun down the tower onto a screen at ground level. Below the tower are instruments such as spectographs and spectroheliographs for analyzing this image.

Another type of solar telescope is the *coronagraph*. It reduces the glare of the photosphere by means of a polished metal disk and nearly perfect lenses. In this way, astronomers can study the corona even



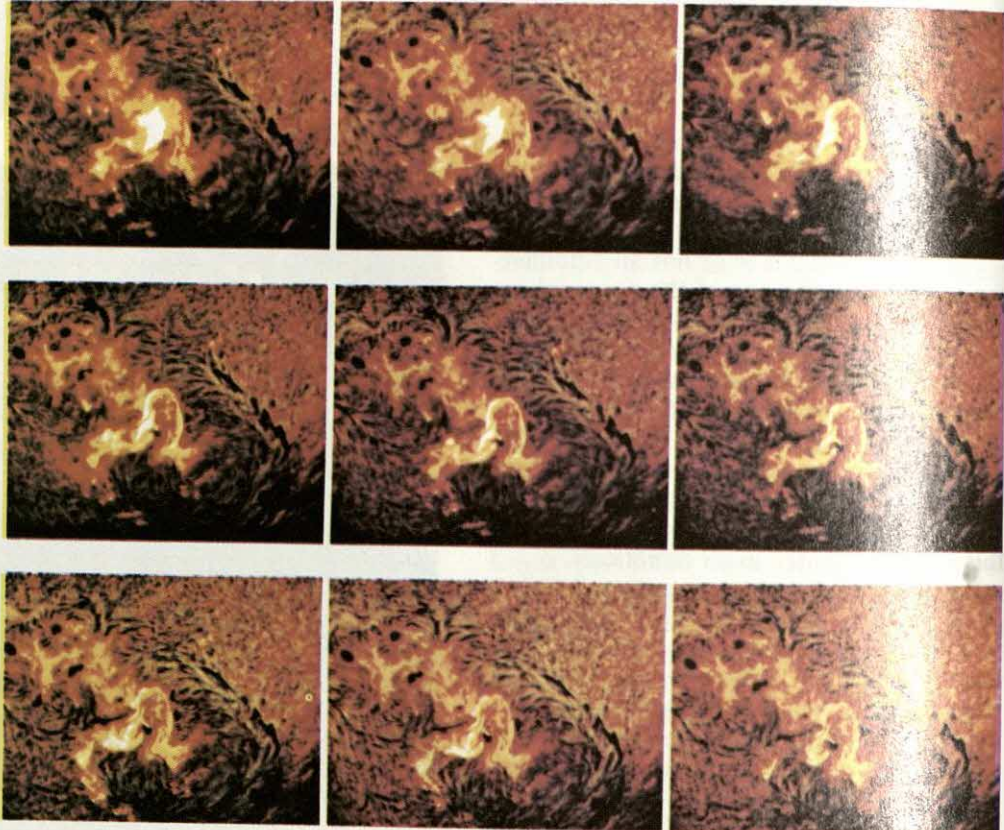
Société astronomique de Suisse

The last phase of a solar eclipse. During an eclipse astronomers can study the chromosphere, the layer of gas close to the surface of the sun. The bright spots along the rim of the sun are solar eruptions occurring in the chromosphere.

when there is no total solar eclipse. The coronagraph is usually located on a mountaintop, where the air is thin and dust-free.

Radio telescopes are used to detect radio waves, which are normally "broadcast" by the chromosphere and the corona. A solar radio telescope is called a *radioheliograph*.

The solar instruments described so far are ground-based—that is, they are located on or very near the surface of the earth. In addition, instruments are being sent into space, aboard small unmanned astronomical observatories. The craft are placed into orbit far above the earth's disturbing atmosphere. Some satellites, such as the U.S. OSO (Orbiting Solar Observatories), are devoted to studying the sun. These vehicles make observations by remote control from ground stations. OSO satellites detect solar ultraviolet radiation, X rays, and particles usually absorbed high in the earth's atmosphere. The U.S. manned space observatory, Skylab, and its Russian counterpart,



Lockheed Solar Laboratory

Solar eruptions send huge amounts of radiation into space. Some of this solar radiation penetrates the earth's magnetosphere.

Soyuz, have also carried out such observations. The U.S.-West German Helios satellites, which have made the closest approaches to the sun, have also gathered much information.

Much routine solar observation is carried out by means of ground-based optical instruments of several types. They analyze sunlight for clues to conditions in the sun. These instruments are commonly used with solar telescopes.

One such instrument is the *spectrograph*, which, like the spectroscope, breaks up light from a telescopic solar image into a series of colored bands from violet to red called a *spectrum*. The temperature of the sun's surface is found, for example, by measuring the quantities of energy in the various spectral colors, which represent different wavelengths of light. The spectrograph also reveals the chemical composition of the sun.

Other instruments produce filmed records of gas movements on the sun and study the visible radiation of the sun. There

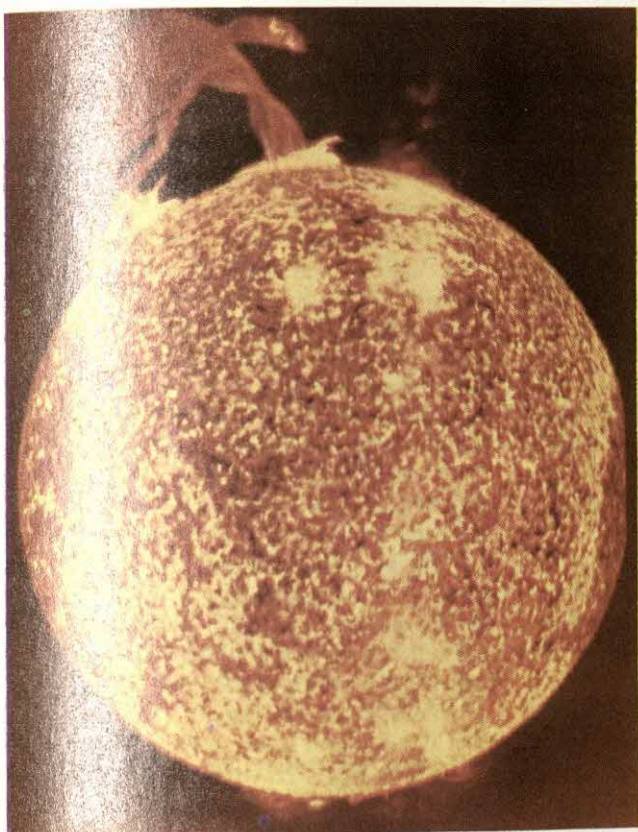
are also devices and films to register invisible solar radiation: X rays, radio waves, and ultraviolet and infrared radiation.

Monochromatic photographs and images of the sun made from otherwise invisible radiation allow us to see aspects of the sun that we would never ordinarily see. Monochromatic photographs, for example, show clearly small details of the prominences, photospheric granulation, flares, and sunspots.

ROTATION OF THE SUN

Like the earth, the sun spins on its axis. The sun is also like the earth in that it spins, or rotates, from west to east. But the sun does not spin at the same speed everywhere. Some parts spin faster than other parts. Astronomers have several ways to calculate the rate of the sun's rotation.

Sunspots and faculae that last for several weeks or months seem to move steadily around the sun. Actually, they are carried along as the sun rotates. The spots and faculae thus enable astronomers to measure



NASA

A spectacular solar prominence spanning more than 588,000 kilometers across the surface of the sun. This Skylab photo also reveals a relative absence of supergranulation near the solar poles.

the speed of solar rotation in those latitudes where spots occur.

In latitudes where spots and faculae are absent, other methods must be used to measure the sun's *period*, or "day"—the time it takes the sun to complete one turn on its axis. An instrument useful for this purpose is the spectroscope. This device breaks up ordinary white sunlight into a spectrum.

The sun's spectrum is crossed by many dark lines. It is really these dark lines that provide clues to solar rotation. When the sun moves away from an observer, the dark lines shift their positions toward the red end of the spectrum. If the sun approaches the observer, the lines shift in the opposite direction, toward the violet end of the spectrum. These shifts of the spectrum lines are called the *Doppler effect*.

As the sun spins on its axis, one side of the sun approaches the observer. At the same time, the other side moves away. From studying the resulting Doppler effect, astronomers find the period of the sun at any latitude.

Investigators have discovered that the

period is shortest at the equator—26.9 days. Rotation becomes slower farther away from the equator. At the north and south poles the period is 34 days. This difference of nearly ten days in period between the solar equator and the poles is due to the fact that the sun is not solid.

CHEMISTRY OF THE SUN

Our knowledge of the chemical elements in the sun is based mostly on study of the solar spectrum. The sun's spectrum, as we said, is covered by many dark lines. These are called *Fraunhofer lines*, after a German physicist, Joseph von Fraunhofer, who discovered them in the 1800's. Fraunhofer lines are also known as *absorption lines*. They represent certain colors, or wavelengths, of light absorbed by different elements in the sun's atmosphere.

The atoms of an element, when hot enough, emit light of certain colors. The atoms also absorb light of these colors, thus producing dark absorption lines. The combination of colors and lines forms a spectrum characteristic of the element. Spectrums of earthly elements have been pro-

duced in laboratories and compared with the solar spectrum. In this way, chemists have learned what elements exist in the sun.

A number of the absorption lines seen in the solar spectrum, however, are caused by certain atoms in the earth's atmosphere. These atoms absorb some of the sunlight passing through the atmosphere. There are about 6,000 of these so-called *telluric* ("earth") lines. But scientists can easily distinguish telluric from solar lines. For example, telluric lines show no Doppler effect from the sun's rotation.

Astronomers have learned that most, if not all, of the chemical elements present on earth also exist in the sun. Hydrogen is the most common solar element, making up more than 80 per cent of the sun's mass.

Helium is second, at 19 per cent. The remaining one per cent of the solar mass consists mostly of the following important elements, in descending order of amounts: oxygen, magnesium, nitrogen, silicon, carbon, sulfur, iron, sodium, calcium, nickel, and a few other trace elements.

The sun is a mixture of gas atoms, atomic nuclei, and still smaller atomic particles. These atomic particles are electrons, protons (positively charged), neutrons (no charge), positrons (positively charged), and neutrinos (no charge).

This entire mass of hot gaseous solar material is called a *plasma*. The high temperatures make it almost impossible for most chemical compounds to exist in the sun.

SOLAR RADIATION

The sun radiates energy at practically all wavelengths. This electromagnetic energy ranges from long radio waves to shorter waves: microwaves and the infrared, light, ultraviolet, and X rays. We see only the light waves. The infrared we sense as heat. The other forms of radiation can be detected only by means of special instruments and films.

There is some question whether the rate of solar radiation is always exactly the same. The amount of light leaving the sun seems to be steady, or constant. But the

quantity of other kinds of radiation emitted by the sun may depend on the number of sunspots present.

The sun also sends subatomic particles into space. Particle emission increases sharply during a sunspot maximum, when solar flares are exceptionally strong. Flares release vast numbers of protons, electrons, and atomic nuclei.

Even during a sunspot minimum, the sun is always emitting particles. A fine "rain" of particles that passes from the corona toward and around the earth is called the *solar wind*.

Many solar particles do not reach the earth or its atmosphere. They are trapped by the earth's magnetic field and become part of a belt of radiation—the Van Allen belt—surrounding the earth.

About 30 per cent of the solar radiation is screened away from the ground by our atmosphere. It is well that this is so, for some of this radiation is deadly.



Strong solar radiation ionizes many of the earth's higher atmospheric gases, producing electrically charged layers. Many scientists call these layers the *ionosphere*. The ionosphere shields the earth below from harmful solar radiation. It also makes long-distance radio communication possible on earth. This is because certain radio waves bounce off the ionosphere back toward the ground instead of going off straight into space. Radio communication on earth may be disrupted from time to time during a sunspot maximum, because intense solar radiation and particles disturb the upper atmospheric layer.

SOURCES OF SOLAR ENERGY

The sun radiates energy at a fantastic rate. Where does all the energy come from? Man has been speculating about this for many centuries.

The most widely held belief until the middle of the nineteenth century was that

NASA and Harvard College Observatory

the sun is a huge ball of fire. If fire were the source of the sun's energy, what would the fuel be? Wood? Coal? Oil? Hydrogen gas? None of these, even in a mass the size of the sun, could ever burn for more than a few thousand years.

Nineteenth-century scientists did not know about nuclear energy. Therefore, if they did not accept the idea that the sun is a mass of flames, they had to look for other causes of the sun's light and heat.

One theory held that solar energy resulted from the impact of meteorites falling into the sun. This idea is not at all sound, since little heating of the sun would result.

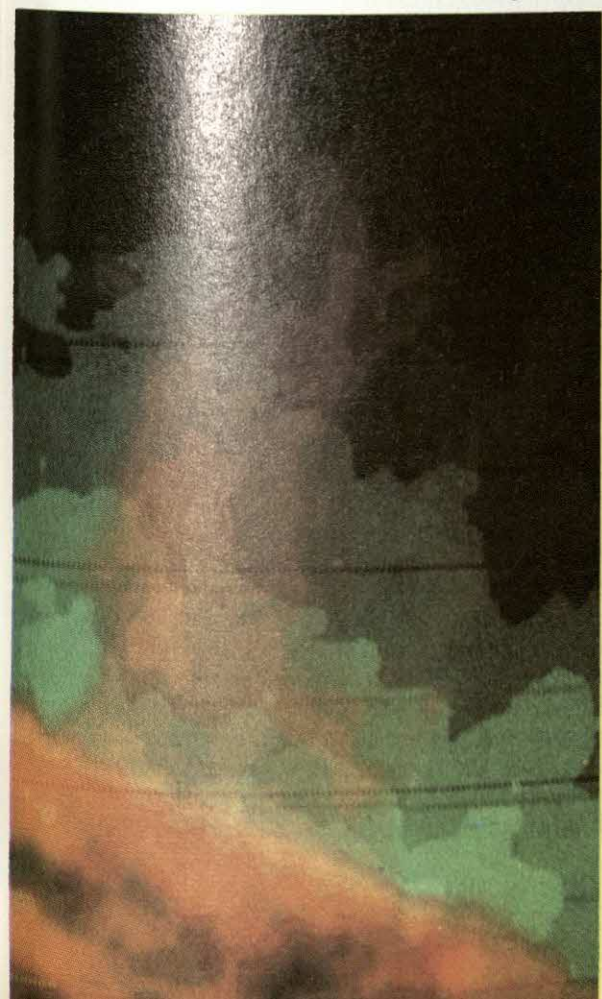
In the 1850's, the German physicist Hermann von Helmholtz offered an explanation of solar energy that satisfied most astronomers. Helmholtz proposed that light and heat energy comes from contraction, or shrinking, of the sun. According to his theory, energy is released as gravity keeps on compressing solar gas into a smaller and smaller volume.

Helmholtz calculated that a decrease in the diameter of the sun of only 85 meters per year would keep up the rate of solar energy output for 25,000,000 years from the time of the sun's origin. In view of the knowledge of Helmholtz and many other scientists of his day concerning the age of the sun, this was long enough. Also, this rate of shrinkage would not have reduced the size of the solar disk noticeably during man's few thousand years of recorded history.

But today, scientists know that the sun is at least 5,000,000,000 years old. Helmholtz's theoretical contraction could never have been in process for all that time.

Another possible source of solar energy was once considered to be radioactivity: the nuclear decay of heavy atoms--a phenomenon discovered in the late nineteenth century. But spectroscopic study of the sun shows that it lacks, perhaps entirely, heavy radioactive elements, such as radium, thorium, and uranium.

Closeup of a limb surge. An explosion is occurring in the region between the chromosphere and the corona, resulting in the release of tremendous amounts of radiation. The different colors indicate different radiation levels.



This Skylab photo shows the sun's hot outer layer, or corona. It has a temperature of more than 1,000,000° Celsius and its characteristic radiation is X rays. Yellow light and temperatures about 6,000° Celsius characterize the photosphere.



NASA

ATOMIC ENERGY OF THE SUN

In 1939, the German-American physicist Hans Bethe proposed an atomic, or nuclear, explanation of the sun's energy. What goes on deep in the sun's interior is probably what happens in a fusion reaction. Four nuclei of hydrogen atoms fuse, or join, to form the nucleus of a helium atom. As a result, terrific energy is released.

Bethe's fusion cycle takes place in six steps, which chiefly involve the element carbon, as well as hydrogen. This complex cycle is thus also called the carbon cycle. The carbon in the sun is used in the process. But carbon atoms are also produced during the reactions. The net result is that the number of carbon atoms is not changed. On the other hand, hydrogen is used up.

We now know that the carbon cycle occurs in very hot stars. But the sun is so cool, compared to other stars, that another reaction is more important: the proton-proton reaction. In this process solar energy comes from the direct fusion of hydrogen nuclei, or protons, without going through the six-step carbon cycle. The proton-proton process also consumes hydrogen.

Scientists have figured that 100 metric tons of hydrogen changing into helium yields more energy than mankind now uses in one year. At the present rate of solar energy production—about 4,000,000 metric tons of mass or matter turning

into pure energy per second—the sun contains enough hydrogen to remain as bright and hot as it is today for the next 30,000,000,000 years and perhaps longer.

SUN AND EARTH

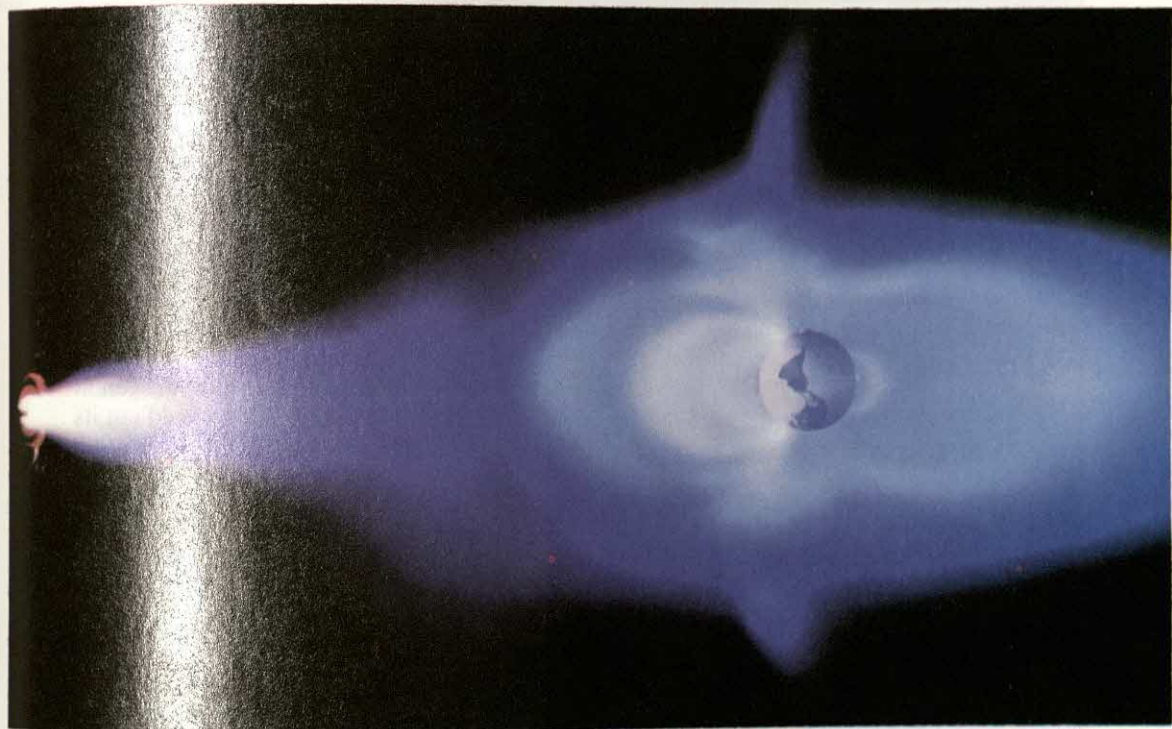
Most of the earth exists as we know it only because of the sun's light and heat. The sun makes all life possible. There would even be no weather without the sun. The sun directly or indirectly provides us with energy to light and heat our homes and to power our machines.

Because it is so small and so distant from the sun, the earth gets only about 1/2,000,000,000 of the total solar energy output. Yet this tiny fraction of the sun's energy that does reach us is awesome in its power and effects.

The quantity of solar energy reaching the edge of the earth's atmosphere equals about 2 calories per square centimeter per minute. A calorie is the quantity of heat needed to raise the temperature of 1 gram of water 1° Celsius.

The rate of 2 calories per square centimeter per minute is called the *solar constant*. About 70 per cent of this energy, on the average, reaches the ground. The earth's atmosphere cuts off the rest.

The sun provides light, warmth, food, and oxygen. Light and warmth are obvious benefits provided by the sun. But how does



NASA

An artist's conception shows the solar wind as a flow of ionized gases from the sun. The wind shapes the earth's magnetosphere into a "tail" pointing away from the sun.

the sun provide food and oxygen? Green plants use solar energy to make food from carbon dioxide and water. As the plants do this, they release oxygen to the environment.

The green plants use the food themselves and, in turn, also nourish many other living organisms, including man. The energy from food is really chemical energy produced from solar energy by plants.

SOLAR ENERGY

Most fuels are a chemical form of solar energy. Coal, for example, is actually the transformed residue of ancient green plants that lived and died many ages ago and were buried in the rocks. When we burn coal, we release solar energy once stored as chemical energy by green plants in their tissues. Petroleum, or oil, is similar to coal in this respect.

Hydroelectric plants, which generate

electricity from running water, also depend on the sun. Without the weather cycle, whose energizing force is the sun's heat, no rain would fall. There would be no running water to move the turbines of hydroelectric power stations.

Solar energy is used more directly in various ways by mankind, but on a very limited scale as yet. Sunlight heats our homes in summertime or the year round in the tropics. Solar heat evaporates seawater in one rather primitive type of salt-making industry.

Solar energy is strong, but so scattered that complex systems of mirrors and lenses are needed to collect it for uses demanding more power than in the examples just given. Once concentrated, however, the power of the sun's heat is fearsome. In a solar furnace, for example, it can melt iron or steel at temperatures of several thousand degrees.

Sunlight generates electricity in light-sensitive cells known as solar batteries. They are used most often in space vehicles and satellites to power their equipment.

Solar engines usually produce steam for power from water heated by the sun. The use of these engines is rather limited at present.

As the earth's fuel reserves are used up, we may have to turn more and more to the sun as a source of energy. Efforts are being made to harness the sun's power for man's use. For a discussion of some of these efforts, see the articles in *The New Book of Popular Science* section on Energy.

PAST AND FUTURE

How long has the sun been radiating energy? How long can it continue to do so? Scientists are attempting to answer these questions. The sun certainly has been shining for at least 5,000,000,000 years. At its present rate, as we stated earlier, it could go on shining for another 30,000,000,000 years at least, if not longer.

But has the sun always been radiating energy at the same rate? Will it continue at the same rate? The answer to the first ques-

tion is "Maybe"; to the second, "Probably not."

At present, the amount of visible solar radiation (light) varies only slightly, if at all. Many astronomers think that this has always been the case. Others, however, disagree. They point to the great ice sheets that from time to time have engulfed much of the earth's surface. These, they say, are evidence that the sun's total radiation may drop off. Even slight decreases would freeze vast areas on earth.

Some astronomers believe that, as the sun grows older, it will use up hydrogen at an increasingly fast rate. This would cut its future life-span to about 10,000,000,000 years. As radiation increases, the sun will become so hot that our oceans will boil away and most life on earth will be killed.

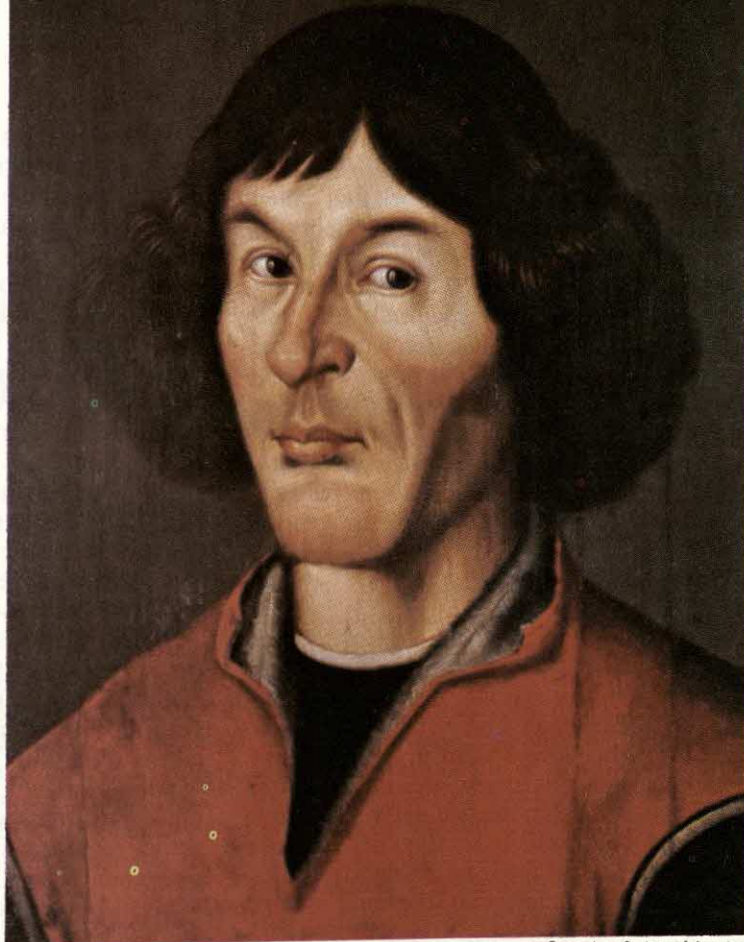
When its hydrogen supply finally gives out, the sun will shrink into a very small star—a so-called *white dwarf*. Later it will die out completely, becoming a dark, cold body. As a livable planet, the earth thus will also die. But since this unhappy event will probably take place, if at all, thousands of millions of years from now, we need not worry.

Solar energy that penetrates the earth's magnetosphere causes the northern lights, observed here in the night sky over Alaska.

Mario Grassi, Geophysical Institute, University of Alaska



Copernicus, the founder of modern astronomy.



Copernicus Society of America

COPERNICUS

by Edward Rosen

Nicholas Copernicus was the founder of modern astronomy.

From time to time in the history of science somebody comes along who starts to question a widely held idea. Such a person must often have a great deal of courage to challenge the accepted theories. He must also have the patience to accumulate the information necessary to show that the accepted idea might be wrong. And finally he must be able to present his new theory in a clear and convincing way. Copernicus was that rare type of scientist.

THE LIFE OF COPERNICUS

His was no easy life. He was born at Torun in present-day Poland on February

19, 1473. His father died when Nicholas was only ten years old, but his mother's brother affectionately looked after the material needs of his talented nephew. The uncle became a bishop in 1489 and arranged that the young Copernicus was elected a canon of the Cathedral of Frombork (or Frauenburg). Canons held their jobs for life and were assured an ample, steady income.

Copernicus acquired the rudiments of astronomy at the Jagiellonian University of Cracow but for his training in canon law transferred to the University of Bologna. There Copernicus had the good luck to come into close contact with a professor who dared to dispute statements made by

Ptolemy, the greatest ancient authority on astronomy and geography. Copernicus also learned of a just-published work that called attention to a grave defect in a theory of Ptolemy. Young Copernicus then began to question whether the ancient Greek astronomer's theory of the structure of the universe was accurate. This idea would occupy the rest of his life. His great work *De revolutionibus orbium coelestium* (or *Revolutions*, as we will refer to it) was written and revised over a period of many years. Copernicus did not consent to its publication until shortly before 1543, the year of his death.

PTOLEMY'S ROTATING SKY

Ptolemy's plan of the cosmos took at face value what we see in the sky. During the day we observe the sun emerge from the eastern horizon, climb steadily upward, and then slip out of sight in the west. At night the heavens reveal the moon, stars, and planets behaving similarly. Ptolemy believed that this daily rotation of the celestial bodies was the real thing. In other words, he believed that the sun, moon,

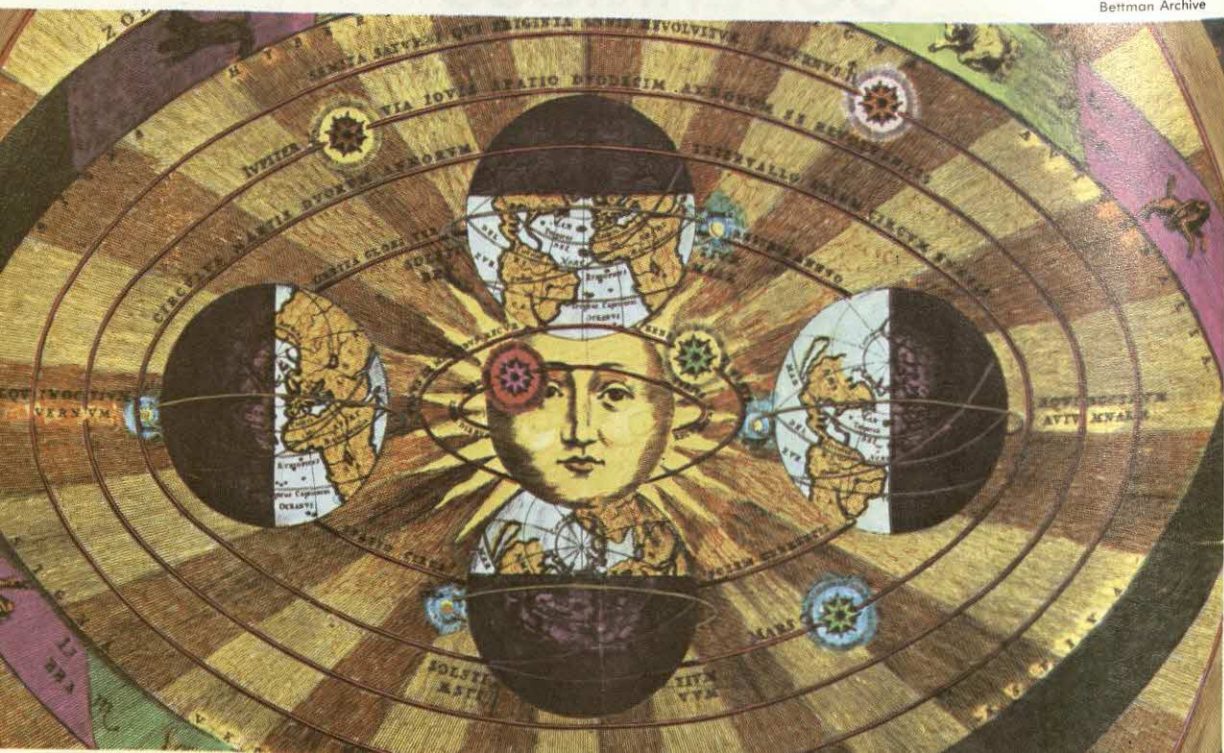
stars, and planets moved around the earth. Copernicus questioned this assumption. He thought that the sun was the center around which the earth and other heavenly bodies turned. He also thought that the earth turned on its own axis.

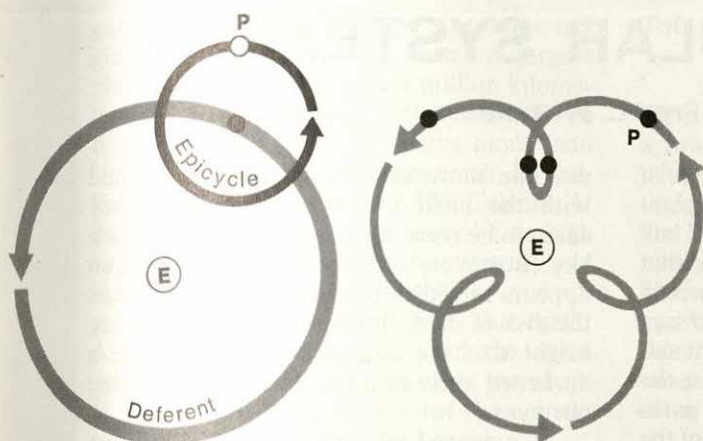
Copernicus was not the first to present this idea. It had been proposed centuries earlier by the ancient Greek astronomer, Aristarchus, but when Copernicus revived the idea he was making a bold break with accepted theories of his time.

Let's examine just one of the explanations Copernicus gave to support his theory and to refute the accepted Ptolemaic theory. An attentive observer of the planets night after night sees them generally follow an eastward course through the pattern of the stars. From time to time, however, a planet seems to slow down, stop, and reverse its direction. This westward or "retrograde" motion goes on for a while, after which the planet seems to resume its normal eastward movement. Ptolemy regarded these regularly recurring loops as real movements and used detailed geometric analysis to try to explain them.

The Copernican universe: the sun is at the center, and the earth and the other planets revolve around the sun. The moon revolves around the earth. This view contradicted the widely held Ptolemaic view in which the earth was at the center and the sun was thought to be between Venus and Mars.

Bettman Archive





Ptolemy explained the retrograde loops of the planets in terms of a complicated system of circular motions upon circular motions. It was assumed by Ptolemy (and by Copernicus too) that celestial bodies move uniformly and in perfect circles. Therefore, explaining motions as irregular as the retrograde loops took some doing. The diagram, far left, shows two of the mechanisms used in Ptolemy's theory of planetary motions. The epicycle rotates at a uniform rate carrying the planet (P) around in a circle. The epicycle in turn is carried around on a deferent that rotates around the unmoving earth (E). The combined motion of epicycle and deferent produces retrograde loops, as seen at near left.

NO—THE EARTH MOVES

It took the genius of a Copernicus to realize that these loops did not really occur. Consider the planet Mars, for example. According to Copernicus' theory, it runs around the sun about once every two years, or in approximately twice the time required for the earth to revolve once around the sun. Start these planets moving at the moment when they are both aligned with the sun. The earth is closer to the sun than Mars is. In a month, the earth on its "inside track" has advanced, say 30°. In about that same time Mars on its "outside track" has moved forward only 15°. Hence the line of sight drawn from the observer on earth past Mars to the more distant stars has moved backward. It appears that Mars has moved westward. But, in fact, Mars has moved eastward but has simply not kept up with the earth's "inside track" movement.

EARTH NO LONGER SPECIAL

In removing the earth from its central position and in making it like other bodies, turning around the sun, Copernicus made his boldest break with accepted ideas. Until then the earth had always been contrasted with the celestial bodies. In this "split-level" concept of the universe, the earth was downstairs, the heavens upstairs. From the innermost of the celestial bodies—the moon—out to the most remote—the stars—stretched heaven. In the Copernican theory, all of this was no longer true. There

was no contrast between heaven and earth. The earth itself was in the heavens along with the other planets. There was nothing unique or special about it. This view disturbed many people, including most of the powerful churchmen of the time. Copernicus hesitated to make his views public, but they gradually became known, and he finally consented to have his work—*Revolutions*—published.

INFLUENCE

When Copernicus' *Revolutions* was finally published, it had a vast influence on the development of science and of thought in general. Since it was published near the time of this death, Copernicus did not suffer any possible repercussions for attacking the established and church-approved view of the universe. But later scientists, who went on to provide the proof of Copernicus' ideas, did suffer at the hands of those who did not want to give up the earth-centered ideas of the universe.

Gradually more and more scientists began to accept the Copernican view and to expand and advance his theories. Johannes Kepler and Isaac Newton, for example, provided an explanation of how the planets moved in their orbits around the sun. It was not, however, until the nineteenth century that the Copernican view of the universe was finally confirmed, and all could accept the statement of a celebrated astronomer that "the work of Copernicus was the greatest step ever taken in astronomy."

THE SOLAR SYSTEM

by Fred L. Whipple

If you were out in space, thousands of millions of kilometers away from our planet, you would see the earth as a tiny ball moving in a wide path around a star that you might recognize as our sun. You would also see, at various distances from the sun, eight other spherical bodies of different sizes—the other planets—all traveling in the same direction in almost circular paths around the sun. Moving around some of the planets you would see smaller balls—the satellites, or moons, of the planets.

In the space between the orbits of two of the planets—Mars and Jupiter—there would be thousands of little planets, or asteroids, also revolving around the sun. Cutting in, this way and that, across the paths of the planets, you would see comets—starry-headed objects, sometimes with long tails streaming after them as they drew near the sun. You might also catch a glimpse of swarms of even smaller particles—the meteors—swirling through space.

All these bodies—sun, planets, satellites, asteroids, and meteors—make up our vast solar system. If you continued to view them for months or for years, you would see that they are moving together through space as a unit, at the rate of some 19 kilometers per second, in the general direction of the blue star Vega.

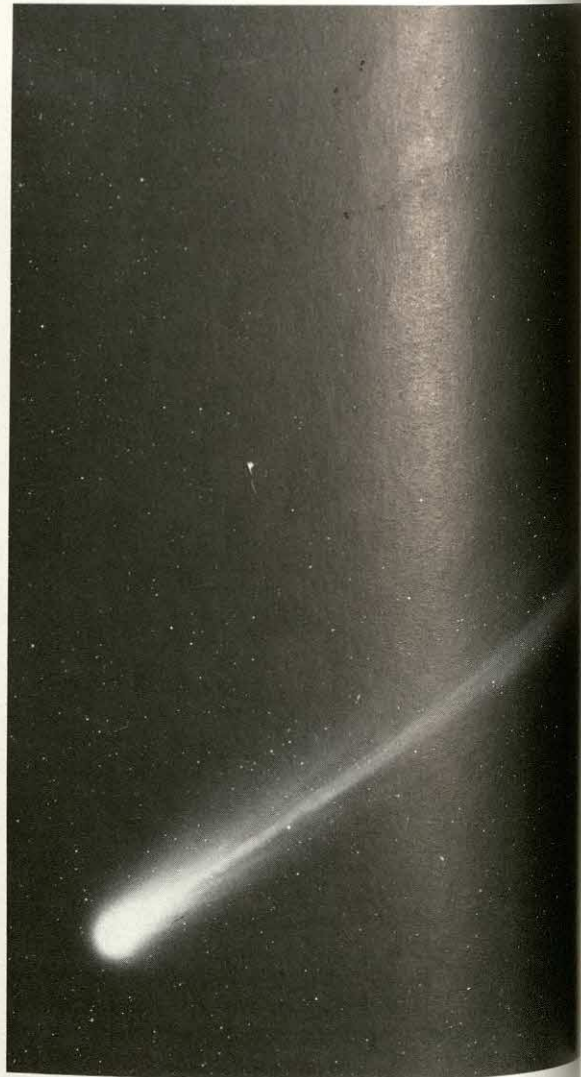
THE SUN

The sun is the very heart of the solar system. It is a typical star—one of the several thousand million in our galaxy. Like the rest, it is an incandescent body made up of highly compressed gases. It is far closer to us than any other star. The distance from the earth to the sun is about 150,000,000 kilometers. The next nearest star, Alpha Centauri, is more than 200,000 times farther out in space. When Alpha Centauri

Comet Tago-Sato-Kosaka. Comets are unusual members of the solar system: they orbit the sun in elongated ellipses and often develop a tail as they near the sun.

and the more distant stars are examined with the most powerful telescopes, they seem to be mere brilliant points in the black sky. But even with the naked eye the sun appears as a disk of about the same size as the disk of the full moon. It is so dazzlingly bright that we must look at it through a darkened glass or a film to avoid damaging our eyes.

Compared with the other stars of our



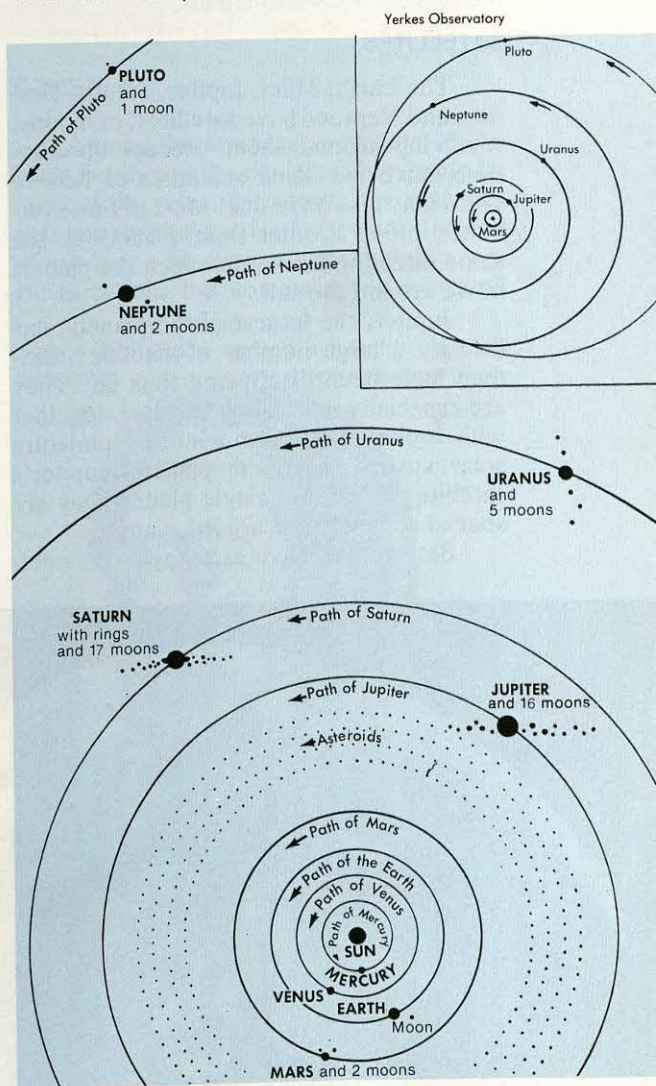
galaxy, the sun is of average size, but it is a giant in comparison with even the largest planets. Its diameter of 1.4 million kilometers is 109 times that of the earth. Even though it is gaseous, it weighs more than 300,000 times as much as the earth. Its surface temperature is about 5,500° Celsius. At its center the temperature may be as high as 15,000,000° Celsius. The heat energy and light energy radiating from the sun make it possible for life to exist upon the earth. Also, without the reflection of the sun's light, we could not see the other members of the solar system, except for the comets and meteors.

Dept. of Astronomy, University of Michigan

THE PLANETS

The planets are the largest bodies in the solar system next to the sun, except for a few satellites that compare in diameter with the small planet Mercury. In order of their distance from the sun, the nine planets are Mercury, Venus, the Earth, Mars, Jupiter, Saturn, Uranus, Neptune, and Pluto. The earth is quite a bit under the average size of the planets. It is much smaller than Jupiter, Saturn, Uranus, and Neptune, which are called giant planets. It is somewhat larger than Mercury, Venus,

The illustrations below show the order of the planets from the sun outward. The larger diagram also shows the satellites of the planets and the asteroids. The smaller diagram indicates the direction in which the outer planets revolve around the sun.



Mars, and Pluto. All the planets occupy no more space in the solar system than nine peas would in a huge football stadium.

Each planet travels around the sun in a giant ellipse, which is very nearly a circle. The orbits of all the planets are more or less in the same plane, except for that of Pluto, the outermost planet. The orbit of Pluto is inclined 17 degrees to that of the earth.

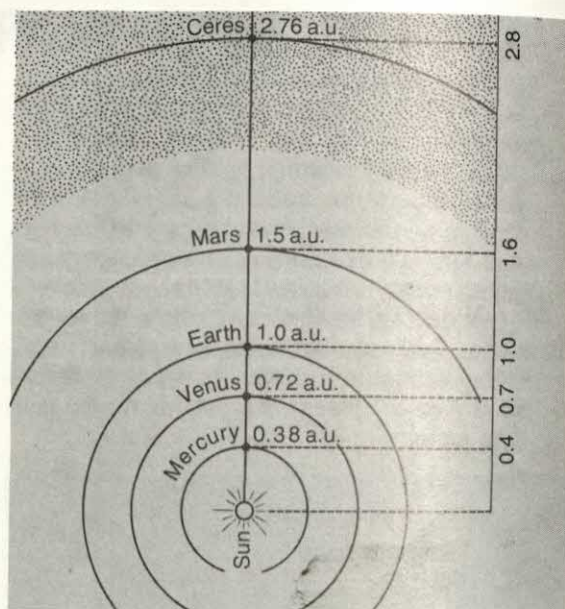
In general, as the distance from the sun increases, the planet paths are more and more widely separated. The 18th-century German astronomer Johann Elert Bode devised a simple rule of thumb, called Bode's Law, for determining the distances of the planets from the sun. Since this law has no real theoretical basis, we need not be surprised that it breaks down in the case of the two most distant planets.

SATELLITES

The Earth, Mars, Jupiter, Saturn, Uranus, and Neptune have satellites, or moons, revolving around them. Recent observations have led some scientists to believe Pluto may also have one. Most of these satellites move around their planets in the same direction as that in which the planets move around the sun.

Jupiter, the largest of the planets, has fittingly a large number of satellites. Sixteen have been discovered thus far. They are especially interesting because, together with Jupiter, they form a sort of miniature solar system. Like the planets, Jupiter's satellites move in a single plane. They are spaced somewhat as are the planets.

Saturn may have as many as 21 satel-

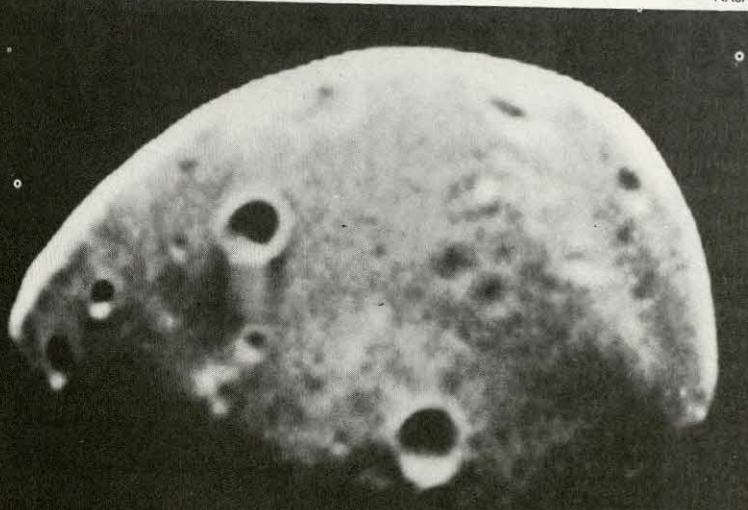


lites, and a vast system of tiny moonlets that reflect sunlight to form a magnificent halo around the planet's disk. Jupiter and Uranus also have rings, much fainter than the bright rings of Saturn. Neptune, too, may prove to be ringed. Uranus has five moons, Neptune two, Earth one, and Mars two. Mercury and Venus have none. Pluto's moon, called Charon, was first detected in 1978. Mars' larger moon, Phobos, revolves around Mars faster than Mars turns on its axis. As a result Phobos rises in the west and sets in the east.

ASTEROIDS

According to Bode's Law, there is a gap between Mars and Jupiter. Astrono-

NASA



Deimos, the smaller of Mars' two satellites, was photographed by the U.S. Viking 1 space probe. Note the heavily cratered surface.

BODE'S LAW OF PLANETARY DISTANCES

Bode's law is a fascinating formula, but it falls short of being a "law." First write 0 and 3, then keep doubling the last number—6, 12, 24, and so on. Add four to each number. Divide by 10 to get the expected distance of a planet from the sun. That distance is in astronomical units (a.u.). One a.u. is the distance from the earth to the sun.

Jupiter
5.2 a.u.

Saturn
9.5 a.u.

In the case of Uranus, Bode's law comes close (19.6 estimated a.u. vs 19.2 actual). But it misses badly on Neptune (38.8 vs 30.1), and on Pluto (77.2 vs 39.5).

5.2

10.0

← BODE'S SEQUENCE →

mers in Bode's day felt sure that there must be an undiscovered planet between these two bodies, and they eagerly searched the skies for it. On the night of January 1, 1801, the Italian astronomer Giuseppi Piazzi discovered a small celestial body, which he took to be a planet, in the space between the orbits of Mars and Jupiter. This body, which was later called Ceres, was found to have a diameter of only 770 kilometers. Other small planetlike bodies were found in the course of time in the gap between Mars and Jupiter.

Today more than one thousand of these small bodies have been discovered, and it is estimated that there are more than 50,000 in all. They are known as minor planets, or asteroids. The orbits of some extend beyond the Mars-Jupiter space. The combined mass of the asteroids is only a fraction of the earth's mass.

Nobody knows how the asteroids originated. According to one theory, they represent the fragments of a big planet whose orbit lay between the orbits of Mars and Jupiter and that broke apart for some unknown reason. This theory has, however, been recently shown to be untenable. More probably, there were several smaller planets in this area, and they collided, breaking up into many tiny asteroids.

COMETS

Among the strangest members of the sun's family are the comets. They abide by none of the rules that govern the nine planets and the thousands of asteroids. Instead of moving in nearly circular orbits in a single direction, the comets revolve around

Johann Elert Bode devised a rule for planetary distances.
New York Public Library Picture Collection



THE PLANETS

| Planet | Mean Diameter in Kilometers | Average Distance from the Sun in Kilometers | Period of Revolution | Period of Rotation | Number of Moons | Atmospheric Composition | Albedo** |
|---------|-----------------------------|---|----------------------|--------------------|-----------------|-----------------------------------|----------|
| Mercury | 4,862 | 58,000,000 | 88 days | 59 days | 0 | He | .06 |
| Venus | 12,190 | 108,000,000 | 225 days | - 243 days* | 0 | CO ₂ | .76 |
| Earth | 12,725 | 149,600,000 | 365 days | 23.9 hours | 1 | N ₂ , O | .36 |
| Mars | 6,780 | 228,000,000 | 1.9 years | 24.6 hours | 2 | CO ₂ | .16 |
| Jupiter | 142,860 | 779,000,000 | 11.9 years | 9.8 hours | 16 | NH ₃ , CH ₄ | .73 |
| Saturn | 120,000 | 1,428,000,000 | 29.5 years | 10.6 hours | 17*** | NH ₃ , CH ₄ | .76 |
| Uranus | 50,100 | 2,875,000,000 | 84 years | 24 hours | 5 | CH ₄ , H ₂ | .93 |
| Neptune | 48,600 | 4,500,000,000 | 164.8 years | 22 hours | 2 | NH ₃ , CH ₄ | .84 |
| Pluto | 2,400 | 6,000,000,000 | 248.4 years | 6.39 days | 1 | CH ₄ | .14 |

* Minus sign indicates rotation opposite to other planets.

** Albedo is the reflecting power of a planet or other nonluminous body. It is expressed as a ratio of the amount of light reflected from the body compared to the amount of light that falls on the body from an outside source. Mercury, for example, reflects 6% of the light that falls on it.

*** 21 discovered but not confirmed.

the sun in exceedingly elongated ellipses and in every conceivable direction. Much of the time they are so far away from the sun that they are invisible even in our largest telescopes. It was formerly thought that some comets approach the sun from far beyond the solar system and that, once they withdraw from the sun, they never return. Today it is generally agreed that comets are members of the sun's family.

When first discovered, comets usually appear as faint, diffused bodies, which are densest at the center. The dense part, which looks like a tiny star, is called the nucleus. The veil-like region that surrounds it is known as the coma. As the comet approaches the sun, the coma becomes brighter. At a distance of some 160 million kilometers from the sun, some comets begin to show nebulous matter streaming away in the direction opposite the sun and forming a tail. This apparently consists of very thin gases that shine by absorbing and reflecting the sunlight that falls upon them. They are forced away from the sun by the pressure of the solar wind. Many comets never de-

velop a tail. They remain vague and diffuse bodies even when they are close to the sun.

METEORS

It is believed that comets may break up into the particles that are sometimes seen entering the earth's atmosphere as meteors. Meteors range in size from fragments no larger than pinheads to huge stones weighing many tons. We become aware of meteors only through the bright light produced when they collide with air molecules in our atmosphere. Most meteors disintegrate after striking the atmosphere. Some of them, however, land on the earth. These are called *meteorites*. A small number of meteorites are thought to be of lunar origin; some may be fragments of asteroids; and some may even have a Martian heritage.

The vast majority of meteors are exceedingly small. The bright flash of light seen when a meteor passes overhead is usually caused by an object the size of a pea or smaller. In the heavens these tiny parti-

cles form great clouds of dust in certain areas of the sky. It is probable that such clouds, reflecting the light of the sun, cause the faint glow in the heavens that may sometimes be seen just before sunrise or just after sunset.

EARLY IDEAS OF THE SYSTEM

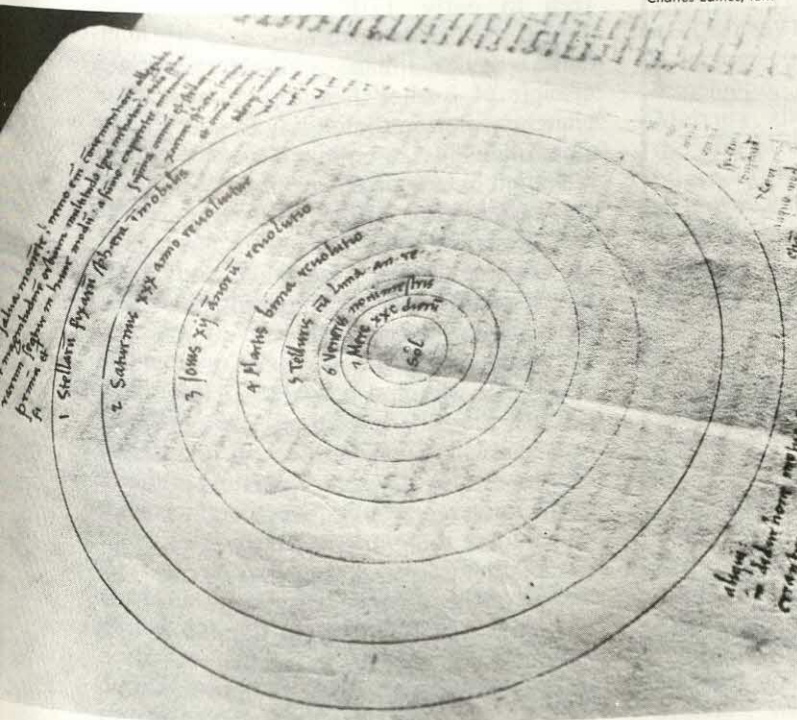
In the light of our present knowledge of the solar system, it is hard to realize that only a few centuries ago men had an entirely erroneous conception of the relations between the sun and its family. A few ancient Greek astronomers suggested that the earth moved around the sun, but this idea won scant support. Most scholars in antiquity held an entirely different sort of theory which to us seems fantastic. They did not realize that the earth itself is a planet, or wanderer in the heavens, which incidentally, is what planet means. The earth, they thought, hung motionless in the very center of the universe. They held that each of the five planets then known (Mercury, Venus, Mars, Jupiter, and Saturn) was attached to a great invisible crystal sphere. The moon and the sun were attached to other spheres.

The crystalline spheres, set one within the other, revolved around the earth, carrying with them the heavenly bodies that were attached to them. This theory satisfied most people; yet careful observers of the sky found in time that it could not explain certain phenomena.

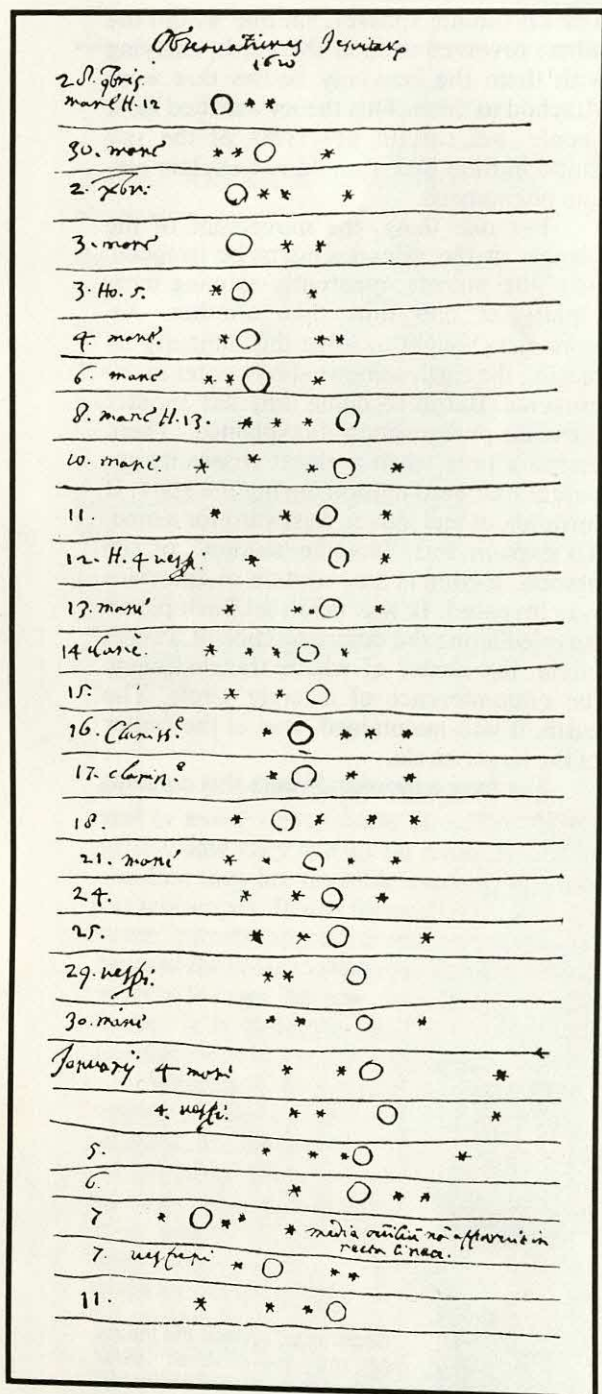
For one thing, the movement of the planets in the skies seems to be irregular, with the planets apparently moving more rapidly at one time than another. Astronomers sought to solve this difficulty by placing the earth somewhat off-center in the universe. But in so doing they left another puzzling phenomenon unexplained. There comes a time when a planet ceases its apparent eastward motion among the stars. It turns about and moves westward for a time. To explain this "reverse motion" of the planets, a complicated system of epicycles was invented. It was held that each planet traveled along the circumference of a small circle, the center of which traveled along the circumference of a larger circle. The earth, it was maintained, was at the center of the larger circle.

For over a thousand years this concep-

Charles Eames, IBM



Copernicus revived the theory of the sun-centered solar system. A portion of his manuscript describes their order relative to the sun.



Galileo's drawings of Jupiter and several of its moons. A sphere represents the planet; asterisks are used to represent moons. At certain times only two of Jupiter's moons were visible to Galileo; at other times he could see three or four.

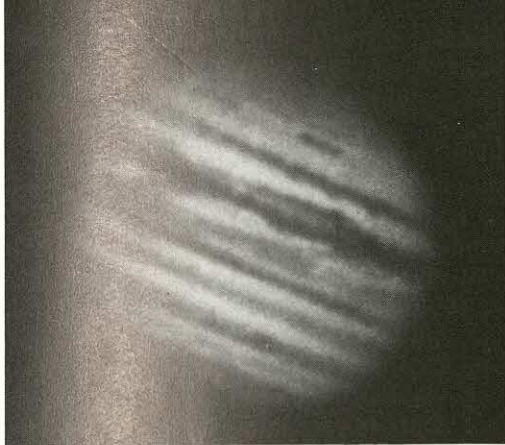
Yerkes Observatory

tion of the universe prevailed. It was taught in the schools of the late Roman Empire and in medieval universities. In the first half of the sixteenth century, however, a Polish astronomer, Nicholas Copernicus, revived the suggestion that the earth moves around the sun. He held that the other planets also revolve around that heavenly body. Copernicus's system, presented to the world in 1543, was called the heliocentric theory since it placed the sun (*helios*, in Greek) at the center of the universe. The Polish astronomer's ideas were quite sound, on the whole, but they were based on insufficient observation. They were inadequate in some respects and wrong in others.

MOTIONS OF THE PLANETS

It required the efforts of several other great astronomers to prove that the heliocentric system of Copernicus had a solid core of truth, for all its defects. The 16th-century Danish nobleman Tycho Brahe made a long and accurate series of observations, which later investigators used as a point of departure. Galileo Galilei, a great Italian physicist, who lived from the mid-16th to early 17th century, staunchly upheld the Copernican system in his influential writings. Johannes Kepler, a German disciple of Brahe, drew up three laws of planetary motion that still hold good today. Copernicus had maintained that the planets move in circular orbits around the sun, and this belief led to confusion. Kepler showed that orbits are ellipses—not circles.

Kepler's laws clearly explained the nature of the planets' movements around the sun, but Kepler did not analyze the force that brings about these movements. This force was first revealed in 1687 when the great English scientist Isaac Newton presented his law of universal gravitation. The law states that every particle of matter in the universe attracts every other particle with a force that varies directly as the product of their masses and inversely as the square of the distance between them. Newton showed mathematically that this is truly a universal law, since it applies not only to objects upon the earth but to heavenly bodies—from meteors to stars—as well.



Lick Observatory



Lick Observatory

The giants of the solar system. Top photo: Jupiter, the largest planet. Lower photo: Saturn, the second largest, with its system of rings.

The law of universal gravitation explains why planets, asteroids, and meteors keep turning around the sun. This huge body binds the planets to itself because of its strong powers of attraction. Universal gravitation also explains why we do not fly off the earth even though, at the equator, it is spinning around at the rate of 1,600 kilometers an hour. People and animals and rocks alike are drawn toward the center of our planet by the force of gravitation. Consequently, the earth has no "bottom side" from which objects can fall off into space. The force of gravitation holds the air and the oceans to the earth. Even the moon, 380,000 kilometers away, feels the effect of the earth's attraction and because of it continues to revolve around our planet.

By utilizing the law of universal gravitation, we can now analyze the motions of the planets with a very high degree of accuracy. We can account for the small deviations that arise as one planet affects the orbit of another. It was the study of

such deviations that led directly to the discovery of the planet Neptune.

After Uranus had been discovered by Sir William Herschel in 1781, careful studies showed that it did not exactly follow the orbit that had been predicted for it in accordance with the law of universal gravitation. This led a young Englishman, John Couch Adams, and a noted French astronomer, Urbain-Jean-Joseph Leverrier, to the conclusion that Uranus was being attracted by another planet even more distant from the sun. Both men calculated the position in the sky of the unknown planet without ever having seen it. On September 23, 1846, on the basis of Leverrier's calculations, the German astronomer Johann Gottfried Galle located Neptune almost exactly where Leverrier had placed it, in the constellation Aquarius. This discovery proved that the law of universal gravitation applies to the universe as well as to objects upon the earth.

Today we realize that this law is not the last word in the analysis of motion in the universe, for it has been modified by the theory of relativity proposed by Albert Einstein. The modification in question is exceedingly slight, however. Thus far it has been applied to only one case of solar-system motion. The perihelion of the planet Mercury, the point where it is nearest to the sun, moves forward about forty seconds of arc farther in a century than is predicted by Newton's law. The theory of relativity predicts very nearly the observed amount. But this exceptional case does not invalidate the law. It is still held to be 99.99999 per cent accurate as far as the solar system is concerned.

We have found out many things about the solar system since the discovery of the law of universal gravitation. We have succeeded in weighing the sun, the earth, and other members of the sun's family, and we have determined the distances between them. Utilizing such devices as the spectroscopic, the spectroheliograph, and the thermocouple, we have analyzed the composition and measured the temperature of the sun and of many other bodies in the solar system.

MERCURY

Mercury is the nearest of the planets to the sun. It is the second smallest but, at certain intervals, one of the brightest. In spite of that fact, it is generally not easy to see with the naked eye. For one thing, it appears in the heavens only during the hours of twilight and dawn, when even very bright stars do not appear at their best. Besides, it is often obscured by haze near the horizon. The great Polish astronomer Nicholas Copernicus once lamented the fact that he had not been able to see Mercury at all in his many years of observation of the heavens. Perhaps this was due to the nature of the district where he lived—the low and misty region of eastern Prussia where the Vistula flows into the Baltic.

MORNING AND EVENING STAR

Mercury makes such a small circuit around the sun that it is always comparatively near that body. It never rises in the morning or sets in the evening much before or after the sun. Because of its appearance sometimes in the east and sometimes in the west, some ancient peoples—including the Egyptians, Hindus, and Greeks—thought of it as two separate heavenly bodies—a morning star and an evening star. The Greeks called the morning star “Apollon” after the god of the sun, and the evening star “Hermes,” the name of the swift messenger of the gods, because the planet’s apparent motion among the stars was so swift. It is said that the Greek philosopher Pythagoras, who lived in the sixth century B.C., was the first to recognize that the morning star and evening star were one and the same heavenly body. That fact was well known to Roman astronomers. Hermes was worshiped by the Romans under the name of Mercury (Mercurius, in Latin).

In the Northern Hemisphere, Mercury

A photomosaic of the heavily cratered surface of Mercury. The photos were taken by Mariner 10 during its flybys of the planet. The largest crater seen here is 200 kilometers in diameter.



can be seen with the naked eye for only a few days at dawn in late summer or early autumn and also at twilight early in the spring. Using telescopes, astronomers do not have to confine their observations of the planet to these periods. When viewed through the telescope, Mercury looks a good deal like the moon as seen with the naked eye. The telescope reveals that the disk of Mercury has phases like those of our moon. It increases from a thin crescent to a full disk and decreases again to a crescent. Then it disappears altogether when the planet is almost directly between the earth and the sun. These phases are not visible to the naked eye.

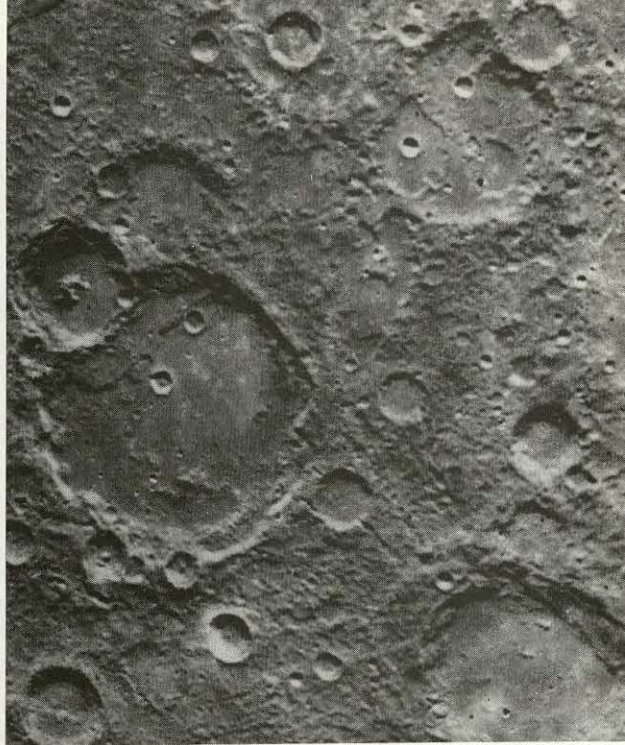
Mercury is almost perfectly spherical. It has a diameter of 4,862 kilometers. Its density—that is, the average amount of mass it contains per unit volume—is very similar to that of the earth. The earth's density is 5.52 grams per cubic centimeter, whereas Mercury's density is 5.44 grams per cubic centimeter. The earth's moon is much less dense.

The average distance of Mercury from the sun is about 58,000,000 kilometers. However, the orbit the planet follows around the sun is shaped like an elongated ellipse. Thus at times the planet is closer to the sun and at other times farther away. In fact, sometimes Mercury is only two thirds as far from the sun as it is at other times.

Mercury races around the sun in almost exactly 88 days. This means that one earth year of 365 days is equal to more than four Mercury years. The speed at which the planet moves in its orbital path varies according to its distance from the sun. When Mercury is farthest from the sun, it travels at 37 kilometers per second. When it is nearest to the sun, however, it speeds up to a velocity of 56 kilometers per second.

HOW LONG A DAY

Like every other planet, Mercury rotates as it revolves around the sun. For a long time, it was believed that Mercury turned on its axis about as rapidly as the earth and so has a day of twenty-four hours. However, the nineteenth-century

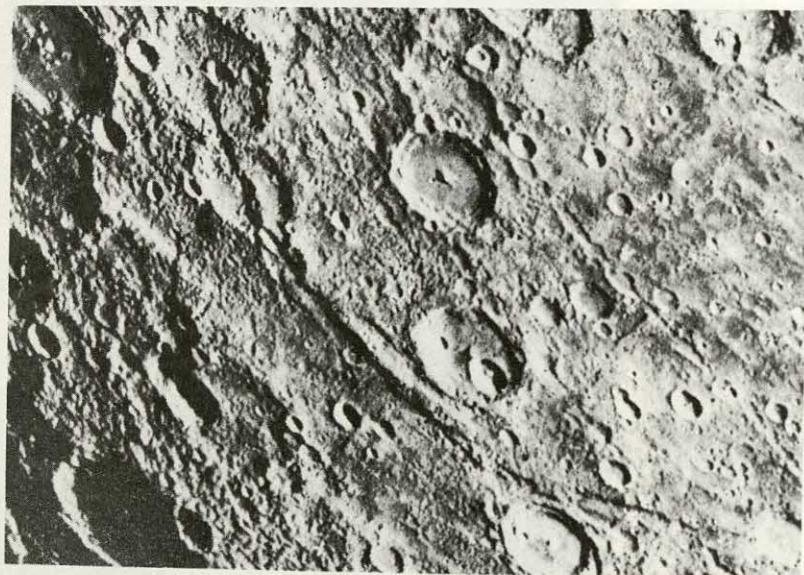


NASA, Jet Propulsion Laboratory

This view of Mercury's south polar region reveals craters nested inside one another (top); large craters with rims peppered with smaller craters (left); and craters with terraced inner walls (left and bottom right).

Italian astronomer G. V. Schiaparelli concluded, from his extensive observations of the planet's markings, that its day equals 88 of our earth days. This idea was held for a long time. Because, supposedly, the period of Mercury's rotation equalled the period of its revolution around the sun (88 earth days), it was thought that one hemisphere of the planet permanently faced the sun, while the other was continually exposed to the darkness and cold of interplanetary space. Supposedly the sunlit side was fiercely hot, while the night side was only a few degrees above absolute zero (-273.16° Celsius).

It is now known that Mercury's period of rotation is not the same as its period of revolution. This was first determined by radar observations in 1965, which showed that the planet completes one rotation on its axis about every 59 earth days. Therefore there is no permanently sunlit side or dark side of Mercury. The entire surface of the planet comes in for its share of the sun's intense heat.



Left: very long cliffs, or scarps, are an unusual kind of surface feature on the planet. They may be formed by compressive forces in the planet's crust. The scarp that extends from upper left to lower right across the picture is more than 300 kilometers long.

Lower left: rays of light-colored material extend out from a relatively new crater (right side of photo), near Mercury's south pole.

NASA, Jet
Propulsion
Laboratory



NASA, Jet Propulsion Laboratory

MARINER 10 FLYBY

Because it is so near the sun, Mercury is very hard to observe and photograph from the earth. Little was really known about the planet's surface until 1974, when man got his first close look at Mercury. A U.S. space probe, Mariner 10, flew to within 700 kilometers of the planet in March of that year. The probe then circled the sun and returned to within about 48,000 kilometers in September. On March 16, 1975, Mariner 10 made its last useful flyby, coming within 320 kilometers of the surface of the planet—much closer than on the previous two flybys. The wealth of photographs and other data collected by special instruments aboard the space probe have revealed much about Mercury.

In the photographs, Mercury is covered with craters and huge basins and looks very much like the earth's moon. Like the moon and Mars, it has a more rugged terrain in one hemisphere and a relatively smooth surface in the other. Temperature measurements indicate that, like the moon, Mercury is probably covered with a finegrained, porous material. But unlike the moon, the earth, or Mars, Mercury appears to have no rifts and cracks in its surface. Instead it is crisscrossed with long cliffs, or scarps. To some scientists this indicates



NASA, Jet Propulsion Laboratory

The dark-rimmed crater at left, surrounded by ejected material, resembles the crater Tycho on the moon. It is 67 kilometers in diameter.

that the planet has shrunk since it was first formed, many hundreds of millions of years ago.

Scientists were surprised to discover that Mercury has a very thin atmosphere consisting of helium. It is so thin that the word "atmosphere" gives the wrong impression, but no such gas envelope had been expected at all. Another surprise was that Mercury has a weak magnetic field. Whether this field is produced by the planet itself or produced in some way by the solar wind—the stream of particles flowing out from the sun—is not yet certain. But at any rate, the interior of Mercury is probably earthlike in composition, with an iron core and a less dense outer crust.

MERCURY IN TRANSIT

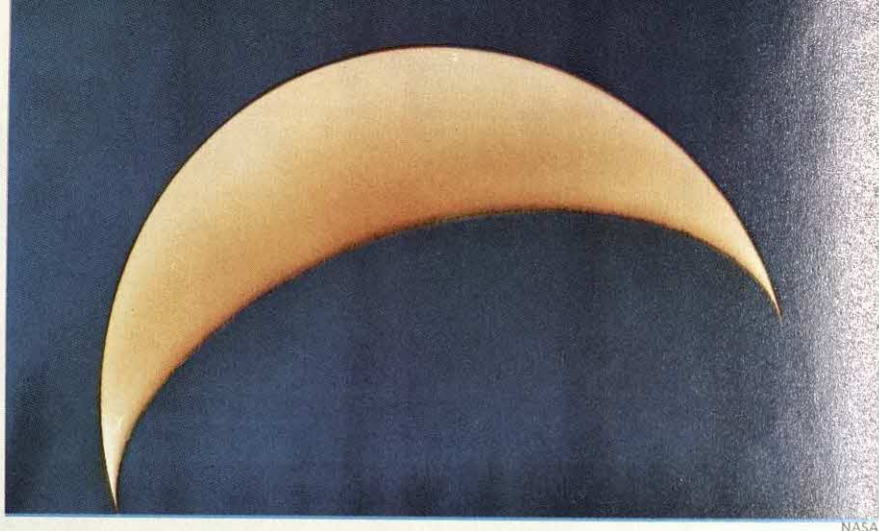
Since the planet Mercury lies within the orbit of the earth, it is possible for it to pass directly between the earth and the sun. Such a passage is known as a transit. If Mercury's orbit were in the same plane as that of the earth, there would be three transits of Mercury each year. It would revolve about four times around the sun in our year and the earth would revolve only once during the same period of time. How-

ever, the orbit of Mercury is inclined by about 7 degrees to the ecliptic—that is, to the plane of the earth's orbit. Mercury, therefore, crosses the ecliptic twice in every 88 days, but generally not in a line with the earth and the sun.

A transit of Mercury can take place only when the planet is between the earth and the sun and is, at the same time, crossing the ecliptic. These conditions are satisfied from time to time, but not at any definite intervals. All of Mercury's transits occur in May and November. Following is a list of the transits from 1970 through 2003:

| | |
|---------------|---------------|
| May 8, 1970 | Nov. 5, 1993 |
| Nov. 9, 1973 | Nov. 15, 1999 |
| Nov. 12, 1986 | May 6, 2003 |

As the planet makes its transit, the astronomer, peering through a telescope, can see the tiny black disk creeping across the dazzling solar background. This cannot be called an eclipse because only an insignificant amount of the sun's surface is obscured. Carefully observing the transit, the astronomer can determine the exact position of Mercury in the heavens and can also obtain added information about its orbit.



Venus shows phases as does the Moon.

VENUS

The beautiful white planet whose orbit lies between those of Mercury and of the Earth is called Venus after the Roman goddess of beauty. The planet is similar to our earth in size and mass. Its diameter is about 12,100 kilometers; the earth's is 12,725 kilometers. Its mass is a little more than four-fifths that of the earth. Its density is about nine-tenths that of our planet.

Like Mercury, Venus is at times an evening star and at other times a morning star, depending on whether it is to the east or west of the sun as viewed from the earth. The planet may rise as much as four hours before the sun and may set as much as four hours after it.

Venus revolves around the sun once every 225 days in an orbit that is very nearly circular. As the planet revolves, it rotates about its axis once every 243.1 earth days, from east to west instead of in the west-to-east direction of most other celestial bodies. The planet is tilted only slightly with respect to the plane of its orbit.

As it proceeds along its orbit, Venus is sometimes on the far side of the sun from the earth, or at superior conjunction. At other times Venus is between the sun and the earth, at inferior conjunction. At superior conjunction it is quite far from earth. But at inferior conjunction it is only about

41,840,000 kilometers away—closer than any other planet. These variations in distance result in notable differences in the apparent size of the planet as viewed from the earth. At inferior conjunction, the apparent diameter is six times greater than at superior conjunction.

In its orbit around the sun, Venus, like Mercury, shows a complete cycle of phases to an observer on earth. These phases are invisible to the naked eye, but may be seen with a small telescope or with good binoculars. When the planet is at the farthest part of its orbit from the earth, it appears as a disk. At inferior conjunction it is almost never visible. About thirty-five days before and after this time, it appears as a crescent and is at its brightest—two and one-half times brighter than when it is seen as a disk.

THICK CLOUD COVER

Venus has been explored by 15 spacecraft of which five were from the United States and ten were from the Soviet Union. Some of these were orbiters, some were landers, and some were both.

The planet is completely covered with opaque clouds, which make an almost perfect reflecting layer. The albedo of Venus is .76; or, to put it differently, the planet reflects 76 per cent of the light it receives

from the sun. By way of comparison, the moon reflects only 7 percent of the light it receives. Because of its high albedo, Venus is the third brightest object in the sky—the sun being the brightest and the moon the next brightest. (The moon appears brighter to us than Venus does because it is so close to us.)

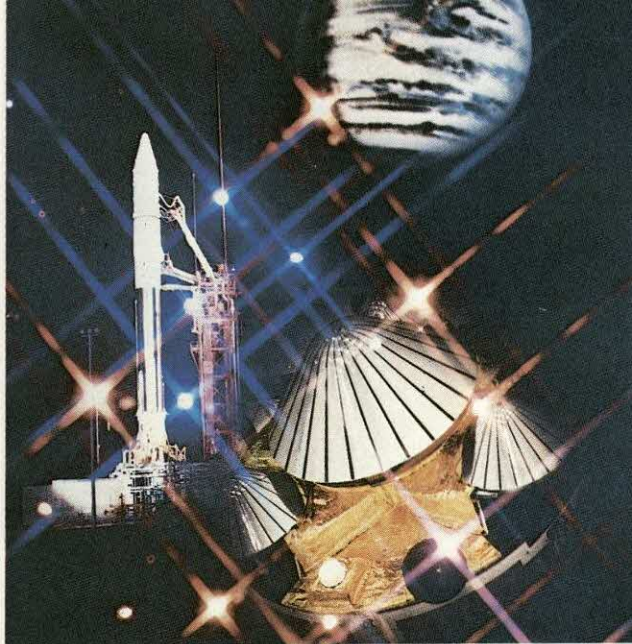
Pioneer Venus and Soviet Venera probes have made photographs of the Venusian surface. Radar was used by the Pioneer Venus orbiter to penetrate the thick veil of clouds and map most of the planet. The mapping reveals mainly rolling plains, with some lowland, two large highlands, and a number of rift valleys. In 1982 Venera 13 and 14 transmitted the first color pictures of the planet's surface. The Veneras' chemical analysis of the Venusian crust showed it to consist of basaltic rock similar to that associated on earth with recent volcanic activity.

HOT, DENSE ATMOSPHERE

Temperatures and pressures on the planet are extremely high. At the surface of Venus, temperatures are about 480° Celsius—hot enough to melt metals such as lead, aluminum, and zinc. These very high temperatures are believed to be due, at least in part, to a process known as the “greenhouse effect.” In this process, the sun's rays pass through the atmosphere and warm the planet's surface. The warm rays are then radiated outward, but cannot penetrate the carbon dioxide of the atmosphere. Instead, they become trapped between the planet's surface and the bottom layer of the cloud cover. The entire process works much like a greenhouse, where a glass or plastic ceiling traps warm air, making a “hothouse.”

Air pressures at the base of Venus's cloud cover are up to 100 times as high as they are at the surface of the earth. This massive pressure is a great obstacle to exploration of Venus. Unmanned probes last scarcely more than one hour under such conditions.

At ground level the atmosphere of Venus is so dense that it is probably very slow-moving. To walk on the surface of the



A composite of the five-probe Multiprobe, its launching silo, and Venus in the background.

planet would be somewhat like walking through a furnace full of boiling oil.

About 95 percent of the atmosphere of Venus is made up of carbon dioxide. Slight traces of oxygen and some hydrogen, nitrogen, neon, and ammonia have also been found in the atmosphere of Venus.

Venus is almost completely dry. Water vapor is present only in very small amounts—less than 1 percent of the total weight of the air. Certainly there is no water on the planet's very hot surface: it would boil away. The solid material that gives the cloud cover its yellowish appearance has been identified as elemental sulfur. Some of the material may be dust blown up from the surface of the planet.

At the equatorial region, hot bands of air (202° Celsius) spiral upward to a height of 70 kilometers. The ascent cools the air to about 13° Celsius, and clouds form. These circle the planet at 360 kilometers per hour and veer poleward. At the polar regions, the air descends and moves toward the equator while it heats up again.

Severe winds and storms probably occur as the air circulates around the planet. One enormous, long-lasting storm, somewhat like Jupiter's Great Red Spot, was observed in the atmosphere. This storm,



The most noticeable cloud formation of Venus, resembling a huge letter "Y", is wrapped around the equatorial area of the equator. This "Y"-shaped formation circles the planet in four to five days. Sometimes the stem disappears, but the arms of the "Y" seem to be permanent.

called the Venusian "Eye," was the size of the United States.

The multiprobe of Pioneer Venus 2 found that there is several hundred times more argon and neon in the Venusian atmosphere than in ours. One theory about the formation of the solar system had predicted that Venus would have less of these gases than the earth. Therefore, this finding has caused some astronomers to rethink this theory.

The findings of all the unmanned probes are that Venus has no appreciable magnetic field, unlike the earth. As a result, there is no zone of trapped radiation in a magnetosphere, such as surrounds our planet. However, Venus does have an ionized atmospheric layer, or ionosphere, above its surface, due to the reaction between its atmosphere with the stream of particles and radiation coming from the sun. This ionosphere is much thinner and lower than the one on earth, but it does protect Venus's surface to some extent from the fierce radiation from the sun.

TRANSITS ACROSS THE SUN

Like Mercury, Venus occasionally transits—that is, passes across the sun's disk, as seen from the earth. Venus comes between the earth and the sun once every

584 days. Transits of Venus, therefore, would occur about once every nineteen months if Venus were always in the plane of the earth's orbit. However, the plane of Venus's orbit is inclined by more than 3 degrees to that of the earth. When Venus comes into inferior conjunction, it is generally either above or below the orbit of the earth. Four times every 243 years, however, Venus is in inferior conjunction and also in the plane of the earth's orbit at one and the same time. Under these conditions a transit takes place.

These transits always occur in June or December, but at irregular intervals. They come in pairs, each transit of a pair being separated from the other by an eight-year interval. After the second of the transits has occurred, there is no other for more than a hundred years. There were transits of Venus in December 1874 and December 1882. The next two will take place in June 2004 and in June 2012.

When Venus is in transit, the unaided eye can make it out as a black dot on the face of the sun. Remember, however, not to look directly at the sun without in some way protecting your eyes. Viewed with the telescope, the planet appears as a black disk, dwarfed by the far larger solar disk. It takes about eight hours or less for Venus to cross the face of the sun. Just as it makes contact with the sun's disk and also just as it leaves the disk, it is surrounded in whole or in part by a bright ring of light.

At one time, the transits of Venus were eagerly awaited by astronomers because they provided a method for calculating the distance of the sun from the earth. A given transit would be viewed from different stations on the earth. It took a longer time for the planet to cross the sun's face when viewed from some of these stations than when viewed from others. Suppose an astronomer knew the length of time of transit as observed from each one of the stations and also the distance between the stations. He would then readily calculate the distance of the sun from our planet. Nowadays, however, much more exact methods have been devised to determine this distance.

THE EARTH

by Franklyn M. Branley

Four planets in the solar system are smaller than the planet earth. Four are considerably larger. The earth in terms of mass, therefore, is not an outstanding member of the vast solar family. However it is the most important of all the planets to you and me, for it is our home in space—the vantage point from which we view the universe.

Many theories have been proposed to explain the origin of the earth and the other planets. They are little more than ingenious conjectures, for they are based on insufficient data. However, though we know little about the beginnings of our planet, we know a great deal about its shape, structure, properties, and motions.

THE SHAPE OF THE EARTH

We know that the earth is round or nearly so. Photographs of the earth taken from rockets, satellites, and other spacecraft far above its surface show distinctly the curvature of our planet. Before this evidence became available, we could infer the roundness of the earth from certain facts. It was known, for example, that the hull of a receding ship, following the curve of the earth, disappears from view before its superstructure does. It was known also that as a total lunar eclipse develops, the earth casts a curved shadow on the moon.

The earth is not a perfect sphere. Technically speaking, it is an oblate spheroid, or flattened sphere. This shape is probably caused by our planet's force of rotation, which deforms the somewhat plastic earth into a form that is in balance with the forces of rotation and gravity. Recent measurements also indicate that the earth is slightly more flattened on the south pole than on the north pole, which makes it slightly pear-shaped.

The diameter of the earth is 12,700 kilometers from pole to pole, and it is 12,750 kilometers along the equator. The difference between the diameters is com-



Official U.S. Navy photograph

paratively slight. If the earth were represented by a globe 18 centimeters in diameter, the polar radius—the straight-line distance from the center of the earth to one of the poles—would only be $\frac{1}{32}$ of a centimeter less than the equatorial radius—the straight-line distance from the center to a point on the earth's equator.

THE EARTH'S MASS AND DENSITY

The mass of an object represents the concentration of matter in it. It is a constant value. In that respect it differs from weight which, as we shall see, is a measure of gravity and changes from place to place.

Various methods have been used to find the mass of the earth. About 1735, the mathematician Pierre Bouguer, while in Ecuador, measured the extent to which a plumb line was deflected by the gravitational pull of a mountain (the peak called Chimborazo). Since he could estimate the mass of the mountain, he was able to estimate the mass of the earth after the deflection was measured.

A sensitive instrument called a torsion balance is generally used in modern times to determine the earth's mass. The attraction of a large ball of known mass for a small ball is compared with the attraction of the earth for the small ball. According to a recent estimate, the mass of the earth is 5,980,000,000,000,000,000,000 metric tons. Or, in other words, the earth's mass is 5.98×10^{27} grams.

To find the density of the earth, we divide the mass in grams by the volume in cubic centimeters. The volume of the earth is 1.083×10^{27} cubic centimeters. If we divide 5.98×10^{27} by 1.083×10^{27} , we get about 5.5 as the figure for the density of the earth. If we mixed together the air, water, and rock of our planet, the mixture would weigh about five and one-half times more than the same quantity of water. The earth is the densest of all the planets.

GRAVITY AND MAGNETISM ON EARTH

Gravity and magnetism are still mysterious forces in many respects. Yet we have gathered considerable information about them. In the seventeenth century, Sir

Isaac Newton clarified our understanding of gravity when he formulated his famous law of gravitation. This states that every particle in the universe attracts every other particle with a force that varies directly as the product of their masses and inversely as the square of the distance between them. This is a statement of universal gravitation.

The term gravity (or more accurately, terrestrial gravity) is applied to the gravitational force exerted by the earth. Gravity is the force that pulls all materials toward the center of the earth. This force becomes smaller as we move away from the center. You are really measuring the force of gravity every time you weigh yourself. If your weight is 65 kilograms, you are pulled toward the center of the earth with a force of 65 kilograms. Your weight decreases as you go farther away from the center. You weigh less on the top of a mountain than in a deep valley; less in an airplane than when on the ground.

In 1600, Sir William Gilbert, an English physician, advanced the idea that the earth behaves like a huge magnet, with north and south poles. This idea is now universally accepted. When you use a compass, the needle falls along the lines of force that run from one magnetic pole to the other. The magnetic poles do not correspond exactly to the geographic poles.

The law of magnets states that like poles repel and that unlike poles attract. Yet the north pole of a compass points toward the north magnetic pole of the earth. This seeming contradiction is due to the fact that in early times when men used natural compasses (lodestones) to direct their ships, they did not understand the behavior of magnets. They noted that one end of the stone pointed toward the north, so they called this the north end. To avoid confusion, we need only give the name "north-seeking pole" to the part of the compass needle that points to the north.

THREE PARTS OF THE EARTH

The earth is made up of three parts: the air, the water, and the solid part, or, as a scientist would say, the atmosphere, hydrosphere, and lithosphere. The scientific

terms are derived from Greek roots: *atmos* means vapor; *hydro*, water; and *lithos*, stone.

The atmosphere. The air surrounding the earth is composed of about 78 per cent nitrogen, 21 per cent oxygen, and one per cent other gases, including water vapor and carbon dioxide. Dust is also present.

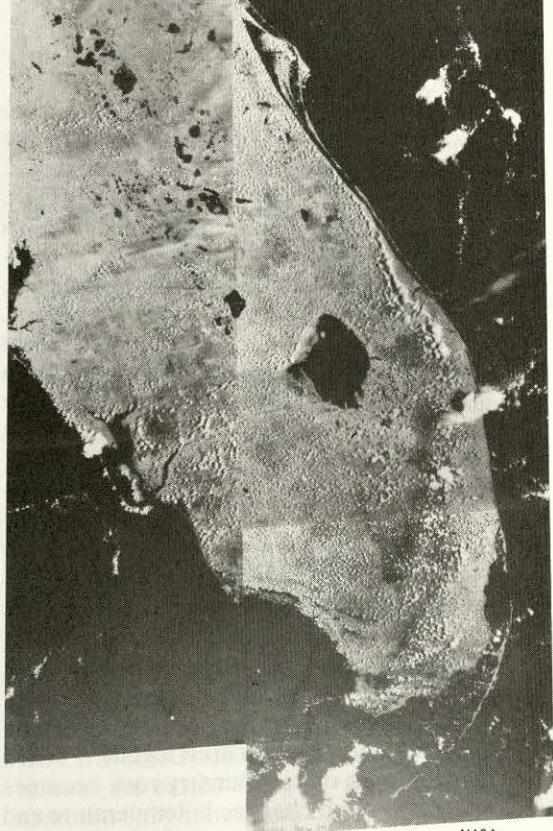
The lower layer of the air envelope is the *troposphere*. *Tropos* means "change" in Greek. The troposphere is the region where great changes take place in temperature, pressure, and water vapor content of the air. It is the part of the atmosphere where our weather occurs. Although most atmospheric changes take place relatively close to the earth, the troposphere extends to an altitude of about 10 kilometers. At the outer limit of the troposphere there is a zone of division between the troposphere and the next sphere. It is called the *tropopause*.

The next atmospheric layer, extending from 10 to about 40 kilometers above the earth's surface, is the *stratosphere*. It is the zone of the strange winds known as jet streams—strong, fast-moving currents of air which may reach velocities of 400 kilometers per hour. Temperature in the stratosphere rises from a low of -60° Celsius at an altitude of 10 kilometers to a high of 0° Celsius at about 40 kilometers.

At this point the stratosphere gives way to the *mesosphere*, which reaches from 40 to about 70 kilometers above our planet's surface. Temperature in the mesosphere ranges from a high of 0° Celsius at 40 kilometers elevation to a low of -90° Celsius at about 75 to 80 kilometers up. The air of the mesosphere is much thinner than that of the stratosphere.

The region of the atmosphere extending from 70 to 400 kilometers above the earth is the *thermosphere*, where the air is extremely thin. Because of exposure to radiation from space and the sun, many of the molecules and atoms are electrically charged; that is, they are ionized. For this reason, scientists often call these layers of ionized air the *ionosphere*.

From 400 kilometers altitude and higher is the *exosphere*, regarded as the



NASA

Space travel has given scientists a new perspective from which to study the earth. Here a composite photo of Florida taken from space. The dark areas are bodies of water, the gray areas tidal marshes, and the white puffs clouds.

outermost fringe of the atmosphere. The extremely thin gas there consists chiefly of hydrogen. The exosphere continues indefinitely into space and eventually merges with the sun's atmosphere.

The hydrosphere. The earth appears to be the only planet that contains large amounts of liquid water. About three fourths of the surface is covered by the oceans. These bodies of water, together with large inland lakes, contribute great amounts of water vapor to the air. They play a large part in the atmospheric changes that we call weather.

Almost 96 per cent by weight of the waters of the earth are made up of hydrogen and oxygen. Sodium, chlorine, and many other elements are also found in oceanic waters. In fact, traces of all the chemical elements would probably be revealed if instruments with enough sensitivity were used.

The plants and animals found in the sea are an immensely valuable resource. They provide man with food, fertilizers, and industrial materials. The ocean is also a vast storehouse of minerals, such as common salt (sodium chloride), magnesium, manganese, gold, iron, copper, uranium, and silver. Some of them, such as salt and magnesium, are obtained from the sea in quantity. Others will no doubt be made available to mankind as more efficient methods for extracting them are developed.

The lithosphere. The solid part of the earth is made up of three types of rock—*igneous*, *sedimentary*, and *metamorphic*—and soil. Soil consists of rock debris combined with organic materials. Igneous rock is derived from the molten, rock-producing matter called *magma*. Sedimentary rock consists principally of rock fragments that have accumulated through untold millennia and that have been pressed together. When igneous rock or sedimentary rock becomes altered through changes in temperature and pressure and other forces within the earth, it gives rise to metamorphic rock.

If we were to dig deep down into the interior of the earth, to the very center of our planet, we do not know positively what we would find. However, earth scientists, gathering evidence from various indirect sources, have developed a more or less

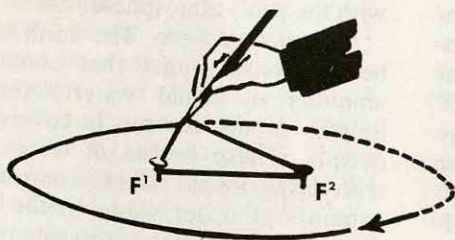
clear picture of the earth's interior.

One of the most distinguished of these scientists, Keith E. Bullen of the University of Sydney in Australia has presented the following view: the earth has a solid outer *mantle* about 2,800 kilometers thick. The crust of the earth makes up only a small part of this mantle. It extends only about 40 kilometers or so below the earth's surface. Beneath the mantle is the *core* of the earth with a radius of some 3,500 kilometers. It is divided into a solid inner core and a molten outer core.

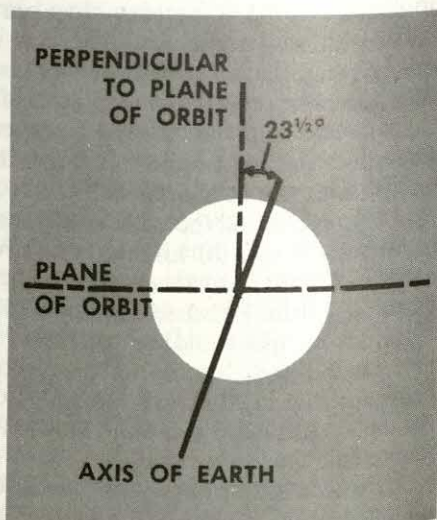
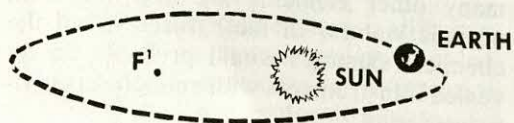
AGE OF THE EARTH

Various methods were used in the past to find out how old the earth is. At one time scientists were sure they could determine the age of our planet by analyzing the salt content of the sea. They took samples of sea water in various parts of the ocean. They learned that about 3.5 per cent of the sea is salt. They argued that sea water was fresh originally, and that it became saltier and saltier in the course of time. By calculating how much salt there is at present in the sea and how much is added each year, they believed they could tell how old the earth is.

We realize now that this method is faulty. For one thing, no salt was added to the water of the ocean during the early



Above: how to draw an ellipse as described in the text. Below: the sun as one of the foci of the ellipse described by the earth in its orbit around the sun. The distance of the earth from the sun varies according to its position in its elliptical orbit.



years of geologic time. Besides, the amount of salt that is added fluctuates from year to year.

The most exact method known to science for determining the age of the earth is based on the study of the radioactivity of certain minerals. In these minerals, one or more chemical elements decay radioactively; that is, their atoms give off very small particles and other radiation. During this process, the radioactive elements are changed into other elements. A given element may also have different forms, or isotopes, which have different atomic weights. Some of these isotopes may be radioactive and undergo change.

Each radioactive decay process takes a fixed length of time, regardless of external circumstances, depending on the isotope and its atomic weight. As the element decays, its quantity in a particular rock or mineral becomes smaller, while the amount of the element it is changing into becomes greater. Knowing the decay times of radioactive elements and the proportions of these elements and of their end products, scientists can calculate the age of the rock or mineral.

There are several radioactive elements, or isotopes, that are commonly used to date ancient objects, including rocks and minerals. These elements are carbon-14 (carbon isotope with atomic weight 14),

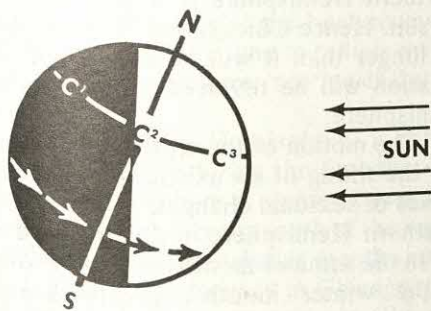
rubidium-87, potassium-40, strontium-90, and uranium-235 and -238. The quantities of isotopes in a sample are measured by radiation detectors and other methods. The age of a rock may be based on one or more isotopes. These methods are not exact, and there may be an uncertainty of as much as several hundred million years. There may not be close agreement in ages measured with different isotopes. The oldest known rock on earth was formed nearly 4,000,000,000 years ago. The earth itself is thought to be about 4,500,000,000 years old.

DAILY MOTION OF THE EARTH

In ancient times and throughout the Middle Ages, many people believed that the earth was motionless. They explained the succession of day and night and the changing position of the stars by saying that the earth stood still and that the sky moved around the earth. We now know that the apparent daily movement of the stars in the heavens is due to the rotation of the earth about its axis. It makes a complete rotation in about 23 hours, 56 minutes, and 4.09 seconds.

One of the best proofs we have of the rotation of our planet is a pendulum experiment first performed in 1851 by a French physicist, Jean-Bernard-Léon Foucault. He suspended a heavy weight at the end of a steel wire, which was suspended from the dome of the Panthéon, a public building in Paris. A pin attached to the end of the weight rested on a circular ridge of sand. Foucault set the pendulum swinging. The pendulum moved to and fro, in the same plane, and the pin at the end of the weight began to trace lines in the sand. As the pendulum continued to swing, the lines followed different directions. There could be only one explanation. The pendulum did not change direction. Therefore it must be the ridge of sand that was turning. Since the sand rested on the floor of the Panthéon and since the Panthéon itself rested on the earth, the earth itself must be rotating.

The device with which Foucault proved the rotation of the earth is called the Foucault pendulum. One of these pendu-



Left: as the earth revolves around the sun, its axis is tilted at an angle of $23\frac{1}{2}^\circ$ from the perpendicular to the plane of its orbit. The motion of the earth around the sun and the tilt of the axis account for seasonal changes on earth. Above: in the summer the Northern Hemisphere is tilted toward the sun and receives the sun's rays more directly. Then any given spot—C—in the Northern Hemisphere will be in sunshine longer than in darkness.

lums has been erected in the General Assembly Building of the United Nations, in New York City.

The alternation of day and night is due to the rotation of the earth about its axis. As our planet turns, a given place on its surface will be in sunlight or in darkness, depending upon whether it is facing the sun or facing the part of the sky on the other side of the earth from the sun.

The earth's rotation causes air currents to be turned toward the right in the Northern Hemisphere and to the left in the Southern.

An interesting effect, due to the rotation of the earth, can be produced by focusing a camera on the North Star and leaving the shutter open for several hours. The stars will appear not as points but as curved lines. This is because the earth, on which the camera rests, has been rotating on its axis.

YEARLY MOTION

At the same time that it rotates, the earth revolves about our star, the sun. It completes a revolution around the sun in 365 days, 6 hours, 9 minutes, and 10 seconds, when reckoned relative to the position of the stars in space. This is called a *sidereal* (star) year.

The orbit of the earth around the sun is an ellipse with the center of the sun at one of the two foci (singular: focus). The definition of an ellipse is that it is the path of a point the sum of whose distances from two fixed points—the foci—is constant. An ellipse can be drawn by the method illustrated on page 100.

Place two pins on a piece of paper resting on a flat surface, as shown. Prepare a piece of string more than twice as long as the distance between the two pins and splice the ends of the string. Place the string around the two pins, and set a pencil in position as indicated. Stretching the string to the fullest extent, pass the pencil point over the paper, going around the pins. The figure drawn by the pencil is an ellipse. The pins will represent the two foci. If this ellipse represented the orbit of the earth around the sun, the sun would be at F^1 or

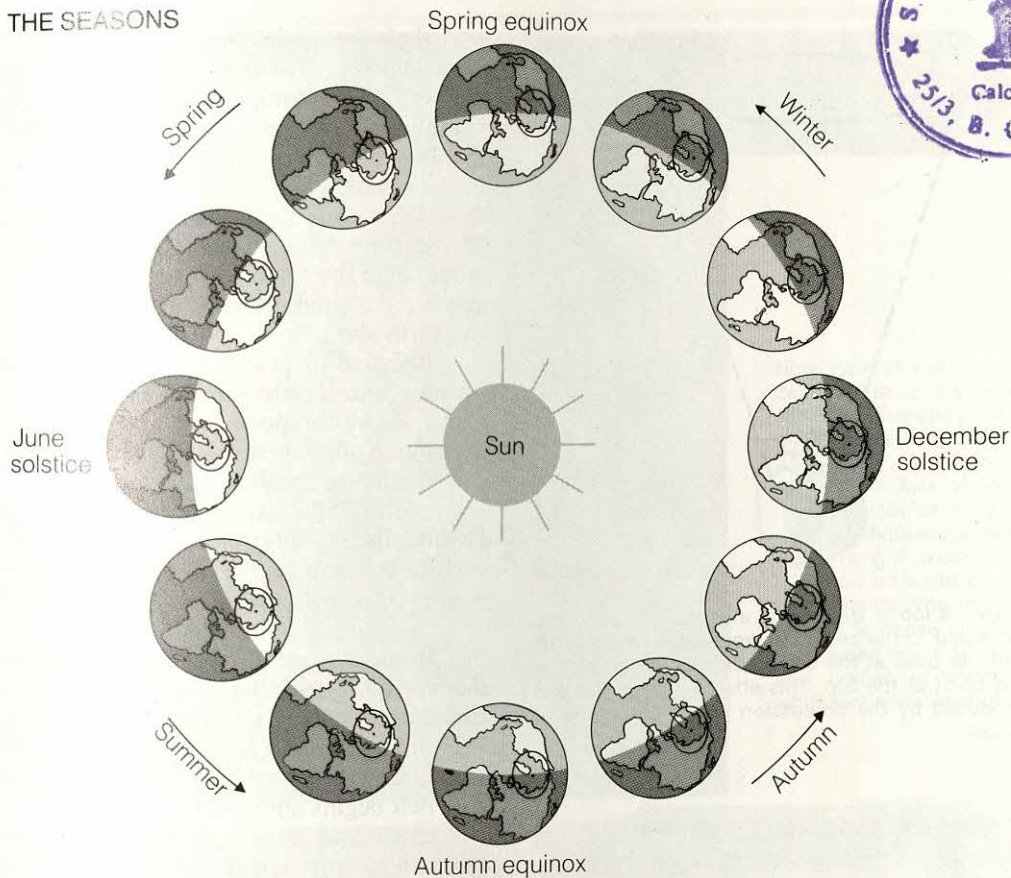
F^2 . Actually the ellipse described by the earth in its movement around the sun is very nearly a circle.

The distance of the earth from the sun will vary according to its position in its elliptical orbit. At perihelion—that is, at its nearest approach to the sun—it is some 4,800,000 kilometers closer to the sun than at aphelion, when it is farthest away. According to a commonly accepted figure, the mean, or average distance is 149,600,000 kilometers. This is often used as a unit of length—the astronomical unit (a.u.)—by astronomers for measuring large distances. For example, instead of giving the distance of the planets from the sun in kilometers, an astronomer could give it in a.u.'s. To Mars it would be 1.5 a.u.'s; to Jupiter, 5.2.

As the earth revolves about the sun, its axis is tilted at an angle of $23\frac{1}{2}$ degrees from the perpendicular to the plane of its orbit. As a result the Northern Hemisphere will be tilted toward the sun in one part of the orbit and away from it in another part; so will the Southern Hemisphere. This accounts for the fact that the days are longer in the summer than in winter in the Northern Hemisphere. In the summer months this hemisphere is tilted toward the sun. Hence, a given spot on the hemisphere—say Chicago—will be in sunlight more than it will be in darkness in the course of a single rotation of the earth. In the winter the Northern Hemisphere is tilted away from the sun. Hence Chicago will be in the shadow longer than it will be in sunlight. The situation will be reversed in the Southern Hemisphere.

The motion of the earth about the sun and the tilting of its axis are the principal causes of seasonal changes. As we saw, the Northern Hemisphere is tilted toward the sun in the summer months and away from it in the winter months. In the summer months, the sun's rays fall more directly upon the earth and heat it more effectively. Besides, as the earth rotates, areas in the Northern Hemisphere remain longer in the sunlight than they do in shadow. That means that for a time more heat is received from the sun than can be radiated away. In the winter months the sun's rays fall more

THE SEASONS



obliquely upon the surface of the Northern Hemisphere. The more the rays slant, the less effectively they heat the surface of the earth. Likewise, because the days are shorter than the nights, the heat received from the sun has more time to radiate away. That is why temperatures are lower during the winter.

The Southern Hemisphere is tilted away from the sun during the time that the Northern Hemisphere is slanted toward it. Hence the winter months south of the equator correspond to the summer months north of it. It is winter in January in Chicago; it is summer in the same month in Buenos Aires.

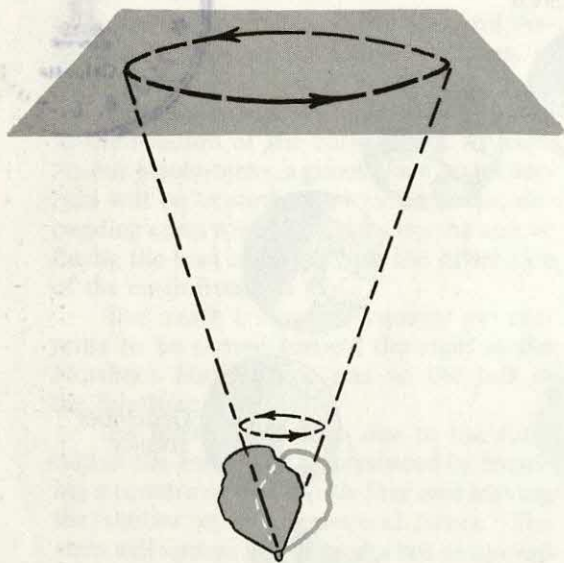
Our planet moves at the rate of about 30 kilometers per second along its path around the sun. This is the average speed. Sometimes the earth moves slower, sometimes faster, in its orbit. It is also traveling through space at a much faster speed as it

follows the sun in its wanderings through the heavens. The sun revolves around the center of the galaxy called the Milky Way, completing a turn in about 200,000,000 years. As a satellite of the sun, the earth takes part in this journey, maintaining a speed estimated at from 190 to 270 kilometers per second.

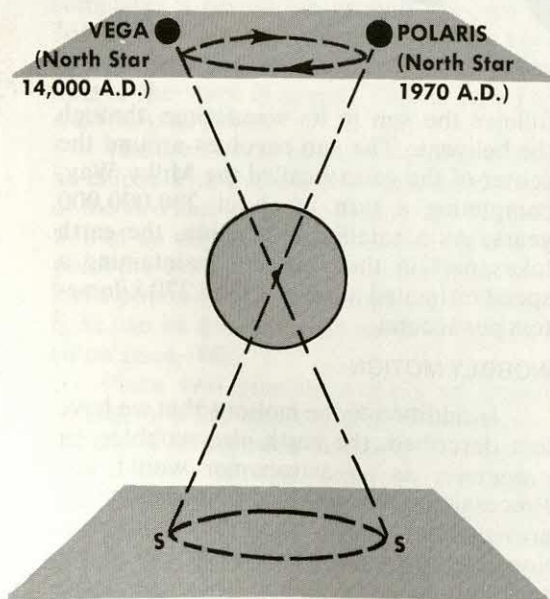
WOBBLY MOTION

In addition to the motions that we have just described, the earth also wobbles, or *precesses*, as an astronomer would say. Precession is due to the combined effects of gravitational attraction and the earth's rotation. The moon (and, to a lesser extent, the sun) is constantly pulling upon the earth. This effect, combined with the earth's rotation, causes the axis of our planet to wobble about its center. As it does so, it traces out two cones in space. These cones have their vertexes, or tips, at the earth's center





Above, a top is spinning at a slant. If its axis were extended to the ceiling, it would describe a cone, with its base at the ceiling and its vertex (tip) at the point of the top. This effect is much like that produced by the precession of the earth, shown below.



The axis of the earth wobbles about its center. As it does so, it traces two cones in space—cones with their vertexes at the center of the earth. Because of this wobbling, called precession, different stars become our North Star in the course of centuries.

and their bases in space above the geographic poles. (See diagram at left.)

We might compare the effect with the spinning of a top at a slant. If we extended the axis of the top, say, to the ceiling of the room in which it is rotating, the axis would describe a cone with its vertex at the point of the top and its base at the ceiling. It would take the top a fraction of a second to complete a single spin. It takes the axis of the earth about 25,800 years.

Because of precession, different stars become our North Star—the one most directly above the north geographic pole—in the course of the years. Right now Polaris is the North Star. Alpha Cephei will be nearest the pole in 7500 A.D.; Vega, in 14,000 A.D. Eventually the full cycle will be completed, and Polaris will be the North Star again.

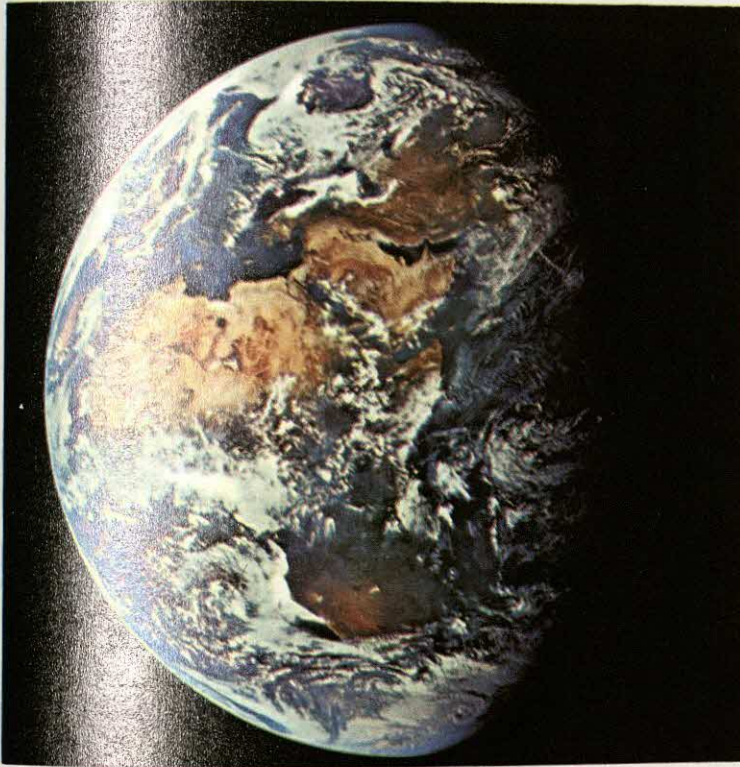
CIRCLED BY A BELT

In the late 1950's it was discovered that the earth is encircled by a large belt of radiation. This belt is now known as the *Van Allen radiation belt*, after its discoverer, the American scientist James A. Van Allen. The belt begins about 650 kilometers above the earth and extends out some 40,000 kilometers into space. It is made up of charged particles radiated by the sun and trapped by the earth's magnetic field.

FUTURE KNOWLEDGE OF THE EARTH

We are beginning to know as much about the earth as a planet as we do about some of the other planets of the solar system. Now that we have ventured into space, we have seen our world in its entirety as a globe. Never before in history have people so extended their observations of the planet earth. Never before in history has the human race been able to see the world in this fashion.

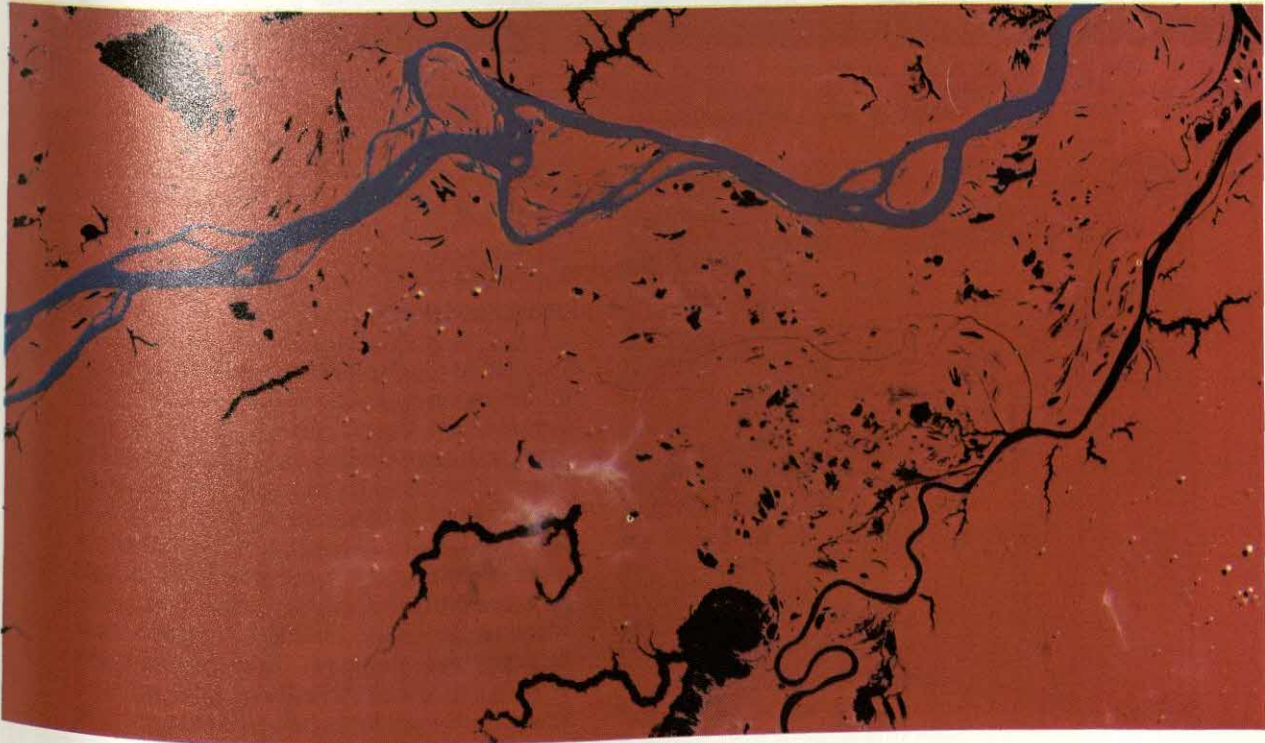
Already scientists have had to change some of their theories about the origin and nature of the earth. They have come to realize how chance has played a role in making the earth such a suitable place for life and intelligence to develop. Despite the general resemblance of the *terrestrial planets*—Mercury, Venus, earth, and Mars—to one other, there are vast differences.



NASA

How part of the earth appears from a distance of 158,000 kilometers. In this photo, taken by the U.S. Apollo 11 crew, northern Africa and Asia Minor are clearly visible. Below: a more detailed photo of the Amazon River (bright blue) region, taken by an earth observation satellite. The red background is dense jungle; the dark areas are river systems and bodies of water, the white wisps, low-level clouds.

General Electric Space Systems





NASA

Photo of the moon taken from space. The landing of astronauts on the moon in 1969 was the realization of a centuries-old dream.

THE MOON

by Cecilia Payne-Gaposchkin
and Katherine Haramundanis

The moon, circling the earth under the pull of gravity, passes across our sky once every 24 hours. Since the dawn of history, people have been fascinated by the moon. They have spun legends about it and have used it in many ways. In prehistoric times, for example, early people observed the phases of the moon and, in some societies, used these phases to form a calendar. In the western world, obvious surface markings on the moon were called "the man in the moon." In China these markings were called "the mortar and pestle and the hare."

When Galileo first turned his telescope on the moon in 1609, he saw its surface in considerable detail. He recognized mountains and large dark areas, which he called "*maria*" because he thought they might be seas. (*Maria* is the Latin word for "seas.") The word "*maria*" is still used, although we know now that there are no visible bodies of water on the moon. As larger telescopes were made, more and more details were noticed and mapped. Beginning in the 1960's, artificial satellites and manned spacecraft were launched from the earth to

pass near the moon or land on it. Pictures obtained by these vehicles have revealed the surface in extraordinary detail, even making visible small boulders about 30 centimeters across.

An earthbound observer of the moon can distinguish light and dark areas. The light areas are generally uplands, and the dark areas low-lying flat regions. Many features are seen to throw shadows on the lunar surface. These shadows can be used to estimate the heights of the features. The telescope also shows bright streaks, radiating from some craters. These *rays* do not throw shadows and are therefore not raised features. Dark, meandering, riverlike features, called *rills*, are probably surface cracks, but no water flows in them. There are also the numerous craters of all sizes. For example, the huge crater Clavius has a 235-kilometer diameter and is surrounded by mountains 5.2 kilometers high.

SIZE OF THE MOON

In order to measure the size of the moon, we must find out how far away it is. This is done with the same methods that are used by surveyors. The angular direction to the moon is measured from two widely separated places on opposite sides of the earth, at the same time. Since we know the distance between the two points of observation, the average distance to the moon can be calculated by simple trigonometry. The average distance is 385,000 kilometers.

The diameter of the moon is 3,480 kilometers, about $\frac{1}{4}$ that of the earth. Other planets in our solar system have satellites, some of which are larger than ours, but our moon is the largest with respect to the size of its parent planet. The largest satellite of Jupiter (Ganymede) is only $\frac{1}{27}$ the size of Jupiter; Titan, the largest satellite of Saturn, is $\frac{1}{25}$ the size of Saturn.

The mass of the moon, which is measured by its gravitational effect on the earth, is $\frac{1}{81}$ of the earth's. Its volume is $\frac{1}{50}$ of the earth's; hence the moon is less dense than the earth. If we take the density of water as 1, then the earth's density is 5.5, or $5\frac{1}{2}$ times that of water. The moon's density is 3.3 on this scale. Actually, the moon is

about as dense as the rocks on the earth's surface and is probably made of similar materials. One possible explanation for the lower overall density of the moon may be its lack of a dense metallic core, such as the earth has.

PHASES OF THE MOON

Anyone who looks at the moon notices that its apparent shape changes from night to night, and runs through a complete cycle in about a month. The changes of shape, the *phases* of the moon, are caused by the changing relative positions of moon, sun, and earth. When the moon is directly in line between the sun and the earth, the sun is shining on the far side of the moon; this phase is known as *new moon*.

As the moon goes around the earth and moves out of the sun-earth line, the illuminated part becomes visible to us as a thin crescent, which increases, or *waxes*, night after night. When the line from earth to moon makes an angle of 90° with the line from earth to sun, we see half the moon's face illuminated. This phase is called *first quarter*. When earth, moon, and sun are again in line, with sun and moon on oppo-

The Taurus-Littrow landing site of Apollo 17 is a narrow, cratered valley. Near this site, unusual orange-colored soil was discovered.

NASA



site sides of the earth, we see the whole face of the moon illuminated. This phase is *full moon*. Thereafter, the moon *wanes*, and the illuminated surface grows smaller. When the direction from earth to moon again makes an angle of 90° with the direction from earth to sun, we again see half the moon's face illuminated in the *last quarter* phase. The crescent continues to wane until new moon is reached again. Between the quarters and new moon, the shape of the illuminated portion of the moon is called a *crescent*. Between the quarters and full moon, the shape of the illuminated disk is described as *gibbous*—that is, not fully circular.

The line that separates the illuminated and dark portions of the moon is known as the *terminator*. The crescent moon, whether waxing or waning, has horns, or *cusps*, which always point away from the sun. Near the time of new moon it is often possible to see the whole disk of the moon faintly illuminated. The light by which we see this phenomenon is sunlight that has been reflected back to the moon from the bright surface of the earth.

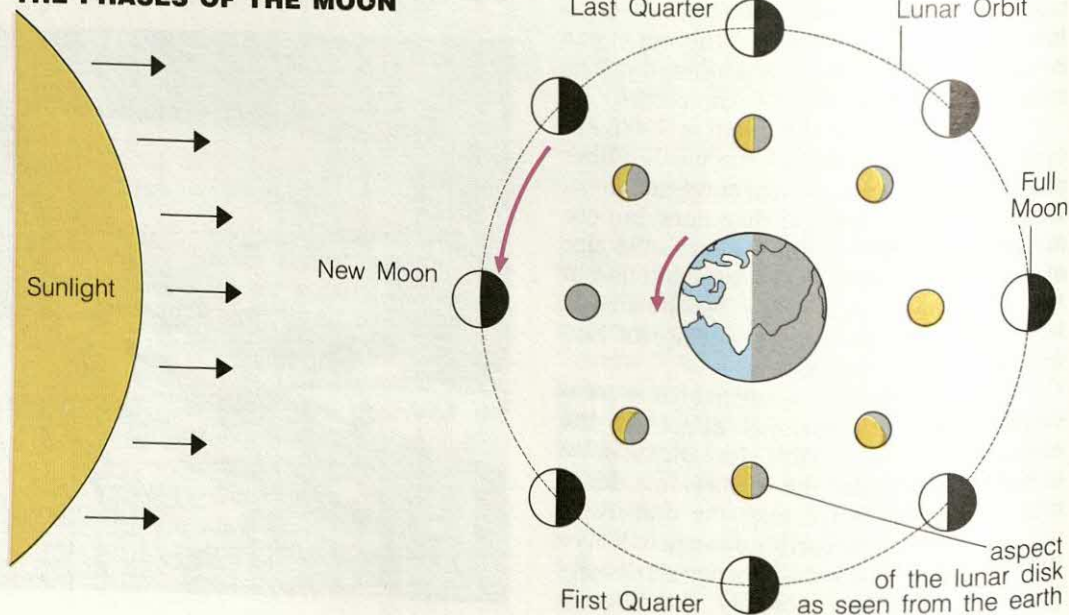
The photographs taken of the various phases show that although the illuminated area changes, we always see virtually the same side of the moon. This is because the moon is gravitationally locked to the earth, and makes one rotation on its own axis in nearly the same time it takes to make one revolution around the earth. However, it should not be forgotten that the sun shines on all sides of the moon in turn during the cycle of phases, while we see only the illuminated portion that faces us.

THE MONTHS: SYNODIC AND SIDEREAL

Several different kinds of month are recognized by astronomers. The simplest one is the *synodic month*, or *lunation*, the interval from one new moon to the next new moon— $29\frac{1}{2}$ days. This, however, is not the time taken by the moon to make one complete orbit around the earth. The moon falls behind because of the earth's motion around the sun, which carries the earth about $\frac{1}{12}$ of the way around its orbit between lunations. The orbital period of the moon, known as the *sidereal month*, is $27\frac{1}{3}$ days; therefore it is nearly two days shorter

The apparent shape of the moon changes from night to night. These changes are caused by the changing relative positions of the sun, earth, and moon.

THE PHASES OF THE MOON





Map of the visible side of the moon with some prominent surface features labelled.

than the synodic month. Because it is gravitationally locked to the earth, the moon rotates on its axis once in a sidereal month.

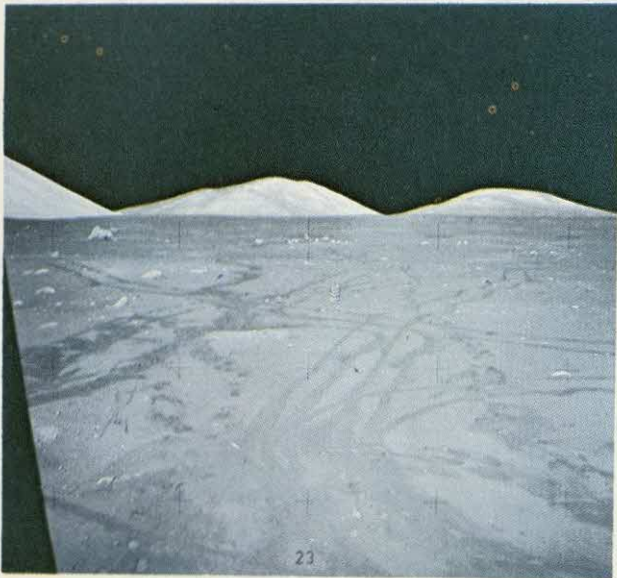
ORBIT OF THE MOON

The motion of the moon is far from simple and has presented a challenge to the ingenuity of astronomers for several centuries. The orbit is not circular. It does not lie always in the same plane. And its shape and position relative to sun and earth are continually changing. For these reasons, the part of the moon seen from the earth varies slightly, so that over a period of time we can view 59 per cent of the moon's surface from a place of observation on the earth. The changes in the moon's orbit run in cycles. Because of this, the visible surface of the moon undergoes rocking motions, or *librations*, which bring small areas near the edges of the observable disk into view. Several decades elapse before all possible areas are visible to viewers on earth. Today, with spacecraft and orbiters, astronomers are obtaining detailed close-up photographs of even the far side of the moon.

ECLIPSES OF THE SUN AND MOON

By a happy accident, the size of the moon as seen from the earth is almost exactly the same as the size of the sun as seen from the earth. This is an extraordinary coincidence, because the sun is about 64 million times the volume of the moon. If the orbit of the earth around the sun and the orbit of the moon around the earth were in exactly the same plane, the moon would pass exactly across the face of the sun at every new moon. Thus there would be an eclipse of the sun once a month. Similarly, at every full moon the shadow of the earth would fall on the moon, and there would be a total eclipse of the moon once a month.

But the moon's orbit is inclined, on the average, by about 5° to that of the earth. This means that the moon can come in front of the sun only near the position where the two orbits intersect, points called the *nodes* of the moon's orbit. When a new moon occurs very near the node, the moon will pass exactly across the face of the sun, and there will be a *total eclipse* of the sun.



NASA

The tracks made by the astronauts who visited the lunar surface will remain unchanged for centuries since erosion occurs very slowly on the moon.

Farther from the node, the moon will cover only part of the sun's face, resulting in a *partial eclipse*. Because neither the moon's nor the earth's orbit is circular, the distances from earth to moon and sun are not constant. As these distances vary, so do the apparent sizes of the sun and moon. At times, the moon may not cover the entire solar disk, allowing a thin rim of sunlight to be visible around its edge—an *annular eclipse* of the sun.

An eclipse of the sun is visible only from the small portion of the earth on which the moon's shadow is cast. The shadow moves rapidly across the surface of the earth, and total eclipses last only a few minutes, as seen from one station. The prediction of the times at which such eclipses will occur involves elaborate calculations which must take into account the complexities of the moon's motion.

An eclipse of the moon takes place when the moon passes through the shadow cast by the earth. They must therefore occur at the time of full moon, when the moon is near the node. There is usually an eclipse of the moon two weeks before or after an

eclipse of the sun. At a total eclipse of the moon, the moon is wholly within the earth's shadow and no sunlight falls on it. When the moon is not wholly within the shadow, there is a partial eclipse. Unlike eclipses of the sun, lunar eclipses are seen from every part of the earth where the moon is visible at the time.

THE TIDES

There are two high tides for every passage of the moon across a meridian, or north-south line, of the earth, principally because of the gravitational pull of the moon on the ocean waters. In the course of a month there are two *spring tides* (when the range of the tide is largest) and two *neap tides* (when the range is smallest). At the spring tides the pulls of sun and moon are in the same line. At the neaps, they make an angle of 90° . Therefore the spring tides occur at full and new moon. The neaps occur at the quarters. In a particular locality the tide may peak some time after the passage of the moon across the local meridian because of the inertia of the ocean waters and the variations of the local coastline.

Tides are not confined to the oceans. They occur in every body of water, in the atmosphere, and also in the earth itself. Earth tides are not so large as the ocean tides, because the earth is essentially a solid elastic mass, whereas the ocean is liquid. Just as the moon raises tides on the earth, the earth raises tides on the moon.

BRIGHTNESS OF THE MOON

The moon has no light of its own, but shines by reflected light. The percentage of light reflected by the moon is known as its *albedo*. On the average the moon reflects only seven per cent of the sunlight that falls vertically upon it. As there are many bright and dark areas on the moon, some areas reflect more light than this, and some less. All the light by which we see the moon comes from the sun, either directly or as *earthshine*, after reflection from the earth. The surface of the earth is a much better reflector than that of the moon. The albedo of the earth can be found by measuring the

brightness of earthshine. It is .33. This means that the earth reflects 33 per cent of the sunlight that falls on it, largely because its atmosphere contains so many clouds. Although the moon is the second brightest object in the sky, it sends us only two millionths as much light as the sun.

MOON TEMPERATURE

When it is midday on the moon, with the sun directly overhead, the temperature is 100° Celsius. At lunar midnight the temperature drops to about -116° Celsius. This difference, so much greater than the temperature extremes known on earth between midday and midnight, is mainly due to the moon's lack of atmosphere. An atmosphere acts as a blanket and prevents excessive cooling and heating. It should also be remembered that the lunar day, or interval between successive lunar midnights, is the time taken by the moon to make a complete rotation relative to the sun. It is therefore equal to $27\frac{1}{3}$ of our days. The lunar surface therefore has longer intervals in which to heat up and cool down than does the sur-

face of the earth in the course of a terrestrial day or night.

Direct evidence that the moon has no atmosphere is obtained by watching an *occultation*—that is, the passing of a star behind the moon. If the moon had an atmosphere, the star would fade gradually, but it always vanishes abruptly and reappears as abruptly at the other edge of the moon.

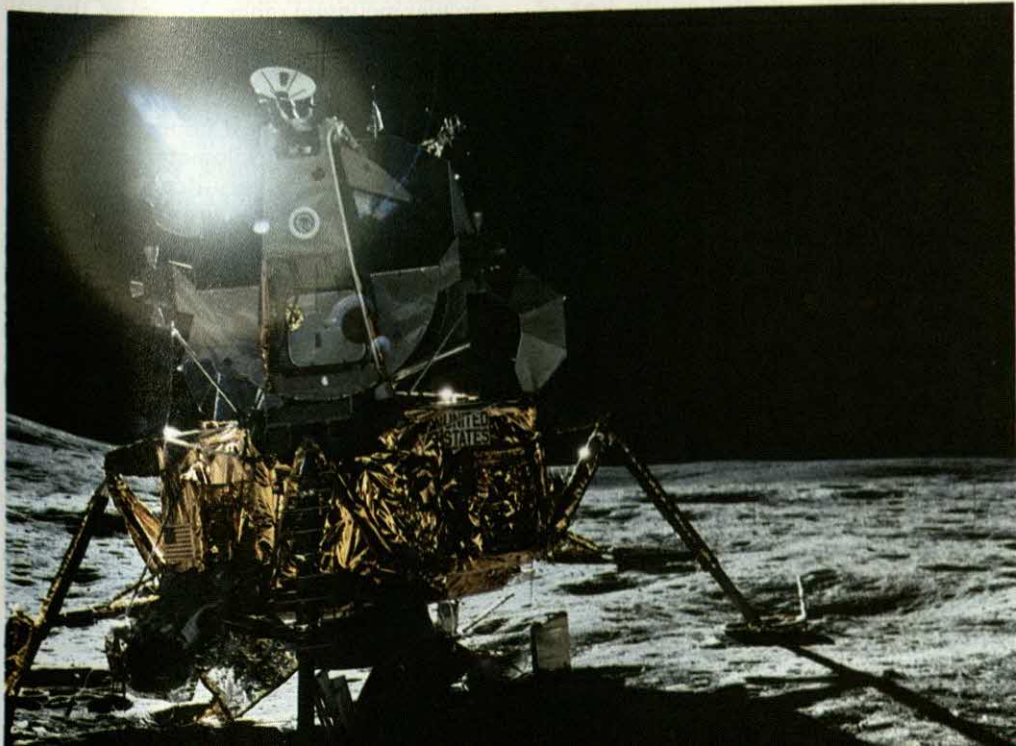
It is not surprising that the moon has no appreciable atmosphere, because gravity at its surface is only one sixth of that at the surface of the earth, and that is not enough to retain most gases at its surface. Small amounts of gas have been seen to exude from certain points on the moon's surface, but they must have been quickly lost into space.

THE MOON'S SURFACE

The largest features on the moon, and the only ones that can be readily seen with the naked eye, are the *maria*. They are darker than the rest of the surface, and are flat areas strewn with small boulders and pocked with craters. Many are surrounded

A lunar module stands on the pockmarked surface of the moon. The surface consists of fine particulate material and rock fragments. This loose cover extends from the surface to depths of about one to three meters.

NASA





NASA

The lunar landscape is dotted with mountains, craters, rills, boulders, and other features. Here Apollo 17 scientist-astronaut Dr. Harrison H. Schmitt stands near a huge boulder that was split by natural forces eons ago.

by mountain ranges. The largest is Mare Imbrium, about 1,100 kilometers across. The Latin names of the *maria*, such as Mare Serenitatis, Mare Fecunditatis, and Mare Tranquillitatis, were assigned in 1651 by the Italian astronomer Giovanni Riccioli. It is remarkable that all the *maria* but one are on the side of the moon facing the earth. Among the *maria*, the most curious is Mare Orientale, situated on the *limb* of the moon (the edge of the moon's disk) and surrounded by three great circular mountain ranges. It is generally thought that the *maria* are hardened lavas that were caused to flow in the far past by the impact of large bodies colliding with the moon, or perhaps by some internal processes.

The moon's surface is very heavily pocked with *craters* of all sizes. They are circular, with raised rims, and some have central peaks. Many craters overlap with other craters, and may occur within the *maria* and on the mountain chains. Some craters appear to be filled with lava, others to be partially buried in dust flows. In the photographs of the great crater Aristarchus, both types of flow may be seen. Aristarchus has a central peak, and its floor seems to be filled with lava. The crater nearby is very shallow and has been nearly

obliterated by a dust flow. A number of much smaller bowl-shaped craters dot the area.

Very large craters are often called "walled plains," because they enclose fairly level surfaces, which may be light in color, like the lunar uplands, or dark, like the *maria*. The crater wall, or rampart, is roughly circular and has often been designated as a circular mountain range. However, the wall is often low in comparison with the surrounding land and even in comparison with the enclosed surface, or crater floor. The diameters of the largest craters reach more than three hundred kilometers. Some craters, however, are very deep, so deep, in fact, that their bottoms are always in shadow. Still other craters are mere pits in the lunar surface, with little or no raised rim surrounding the central depression.

On the earth there are two major types of crater: the volcano and the impact crater. Volcanic craters have steep sides and often a central rise or knob. Crater Lake in Oregon is a good example. Impact craters, such as the Meteor Crater in Arizona and the Chubb Crater in Canada, are shallower relative to their diameters and do not usually possess central peaks. Impact craters can be very much larger than volcanic craters:

the southeastern shore of Hudson Bay may be the remainder of a great impact crater.

Both types of crater can probably be recognized on the moon. There is little doubt that the moon's surface has been heavily bombarded by meteoritic bodies over a long period of time. This bombardment has probably produced not only a large percentage of the craters but also the *maria*. The great dust splash extending halfway around the moon from the huge crater Tycho is clear evidence that it is an impact crater. A large impact crater usually has many small impact craters near it, which were probably formed by the debris thrown out by the original impact. The boulders found in the vicinity of impact craters have a similar origin.

A possible example of a volcanic crater on the moon is Aristarchus. Many small changes have been recorded near it, and quite possibly gases emerge from it on occasion. It has a pronounced central peak and what appears to be a lava-filled floor.

Lunar *rays* are lunar features that extend radially outward from certain craters such as Tycho and Copernicus. They appear lighter than the surrounding terrain principally because they reflect light better. This is probably because they are made of very finely divided particles. The reflectivity of particles depends largely on their average size, finer particles reflecting light more brilliantly under vertical illumination than coarse ones.

Rills are narrow, riverlike valleys made visible by shadows cast into them. They may be cracks in the surface, or possibly canyons formed by ash flow from volcanoes. Some are twisted and tortuous, and some are associated with chains of craters. Some of the straight rills seem to be associated with settling phenomena around the *maria*. Others may have been furrowed by boulders rolling downhill.

The *mountains* of the moon form large, rugged chains principally concentrated around the *maria*. Their heights can be measured by the length of the shadows which they cast. The highest, the Leibnitz Mountains, attain a height of 7.9 kilometers. In proportion to the diameter of its

parent body, this is really higher than Mt. Everest on earth, 8.8 kilometers. Heights on the moon are measured relative to a nearby low-lying area on its surface, whereas heights on earth are measured with respect to sea level, a height about halfway between highest and lowest points on the earth's surface. Lunar mountains seem to have been thrust up as a result of impact rather than formed by folding of rocks as has been the case with many mountains on the earth.

A mountainous feature known as the Straight Wall, in Mare Nubium, south of the lunar equator, appears to be a 110 kilometer-long line of cliffs produced by faulting. A gigantic crack in the lunar surface has caused the development of a cliff face about one-quarter of a kilometer high.

NAMING MOON FEATURES

When the telescope was first turned on the moon, the principal *maria* and mountain chains were named by early observers. The *maria* were given fanciful Latin names, and many mountain chains were called after well-known mountains of Europe. Since that time, advances in observational techniques have revealed more and more detail, and thousands of features have been named. The International Astronomical Union now has the responsibility for assigning names, a task that has been enormously increased by the recent observations of the far side of the moon. Major craters have been named most usually after astronomers or scientists, but other prominent individuals, such as Jules Verne, a 19th-century author, and Plato, a Greek philosopher, have also been commemorated.

ORIGIN AND HISTORY

The history of the moon is bound up with the history of the whole solar system, which is believed, from several independent estimates, to be over 4,500,000,000 years old. There are two principal theories concerning the origin of the moon. The first considers that the moon was once part of the earth and was separated from the earth either by tidal forces or by the gravitational

attraction of a passing star. It was drawn out from that gap in the earth's crust that is now filled by the Pacific Ocean, according to one view. Recent evidence, however, tends to discount the view that the Pacific Ocean basin was the moon's birthplace. It seems the moon has been the earth's companion for many hundreds of millions and perhaps thousands of millions of years. At such an early time, there was no Pacific Ocean as we understand it today.

The second theory suggests that the earth and the moon were formed at about the same time from the agglomeration of cold material then circulating around the sun. Similar processes are suggested for the origins of the other planets and their satellites. In the first theory, it is assumed that the original material was hot, and formed bodies that are now cooling. The second, on the other hand, pictures the formation of the earth, the moon, and the other bodies of the solar system from cold materials, some of which are now heating up as a consequence of internal pressure and radioactivity. A third theory states that the moon is a former planet that was captured by the earth.

According to the second hypothesis, the moon is about as old as the earth—about 4,500,000,000 years. But no doubt the lunar features vary in age, just as earth features do. That the moon at one time during its history underwent extensive development can be seen by a mere glance at its face. Close observation of the moon since the invention of the telescope has revealed very few signs of anything much

happening there in recent times. However, signs of volcanic activity and perhaps radioactive glows seem to indicate that the moon is not really such a dead world after all. The present topography, plus the damage that has apparently been inflicted on its features, certainly means that a lot did happen to the moon in the past.

If the moon once did have an atmosphere and running water, as some authorities believe, then a history somewhat like the earth's must have taken place until air and water vanished. But even with air and water gone, erosion is still possible. Alternate extreme heating and chilling of rock will crack and flake it. Collisions with asteroids and meteorites pulverize the surface and perhaps release volcanic forces. Movements and sliding of rock masses down slopes will cause damage. Even the strong radiation from space may disrupt the lunar crust. There is certainly plenty of rubble, as satellite investigations show. The accumulated ravages of countless centuries of small-scale erosion look impressive. So much so, that persons are tempted to attribute them to tremendous catastrophes, as was the fashion with regard to spectacular earth features in the early days of geologic study. Nevertheless, because of the slow pace of erosion on the moon, experts tend to agree that the moon preserves many ancient features that have long since vanished on the earth. Therefore a trip to the moon may be something like a journey into the earth's past—or even its future.

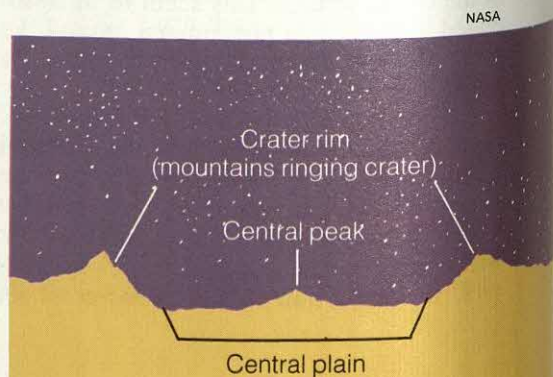
EXPLORATION OF THE MOON

Man has explored the moon at long range, with telescopes and other instru-

Below: the crater Langrenus, with a central core and terraced walls. Below right: cross section of a typical crater. Mountains on its rim slope gently on the outside, steeply on the inside. Smaller mountains may rise from the flat central plain.



NASA



ments. He has also landed unmanned vehicles on the moon's surface and later set foot on it himself. In 1959, the Soviets were the first to land an unmanned craft on the lunar surface. Another vehicle circled the moon and sent to earth pictures of its far side. In the 1960's, the United States photographed the lunar surface in great detail through its Ranger, Surveyor, and Lunar Orbiter satellites. The Rangers crash-landed on the moon. The Surveyors made soft landings and also probed the lunar soil. The Orbiters circled the moon, sending to earth many photographs of both its near and its far sides.

Soviet research has continued with automated sampling and return to earth of lunar surface material, and exploration of the moon with an unmanned wheeled vehicle controlled from the earth. Beginning in 1969, United States Apollo astronauts explored the moon, set up experiments, and returned with lunar rocks.

CONDITIONS FOR THE LUNAR EXPLORER

The view of the moon, as described by Apollo astronauts, gives us an idea of what the lunar explorers encounter. The moon presents a forbidding landscape in various shades of gray. No color enlivens the landscape; even the sky is black. Without an atmosphere there can be no wind, no rain, and no weather. There can be no sound, and no life. In the sunshine the temperature is that of boiling water, but without the blanket of an atmosphere, a step into the shadow brings one to a temperature far below freezing. There are no beautiful sky colors at sunset and sunrise, and no twilight. Like the blue sky, these earthly beauties depend on our atmosphere, and on an airless body they are simply absent.

For these reasons a lunar explorer needs protection, not only against lack of air, but against extremes of temperature. Equally important, he must be protected against the incessant bombardment by cosmic and other rays and particles, and the effects of ultraviolet radiation. On the earth we are shielded from the deleterious effects of strong solar and cosmic rays by our at-

mosphere and the earth's magnetic field. But the lunar explorer has no such natural protection: the moon is without an atmosphere and has an extremely weak magnetic field.

Since gravity at the moon's surface is one-sixth that on the earth, all weights on the moon are diminished by a factor of six. A man who weighed 80 kilograms on earth would weigh about 13 kilograms on the moon, and with the same muscular effort would be able to jump six times as high or lift six times as great a weight as on earth. This compensates to some extent for the weight of the equipment that he has to carry to protect himself from the severe conditions. Similarly the launching of a rocket from the surface of the moon would require only one-sixth of the thrust required to launch the same rocket from the earth's surface. This makes the return journey of an astronaut from the moon and its vicinity easier.

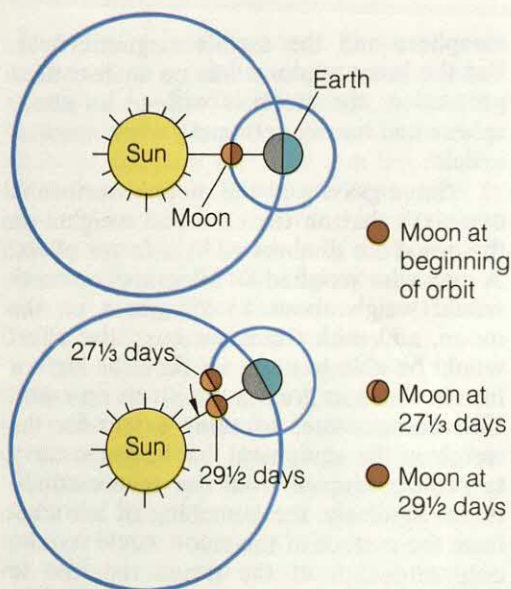
In some cases, his diminished weight works to the disadvantage of the lunar explorer. This occurs if he uses his weight to operate a lever or use a shovel.

Complete weightlessness is experienced by astronauts during space flight, and techniques have been developed to combat the problems involved. All body functions are conditioned by our normal gravitational environment, as are the operations of many instruments that must be used in such explorations.

EARTH AS SEEN FROM THE MOON

A man standing on the moon sees the earth as a disk two and one-half times the size of the moon seen from the earth. Because of its high albedo, the surface of the earth has five times as much reflecting power as the surface of the moon. Also, because of its greater apparent size, the full earth sends about 30 times as much light to the lunar observer as the full moon sends to us. To him our clouds look brilliantly white, the oceans dark blue, and the continents almost uniformly purplish brown.

Because the moon and earth are gravitationally locked, a lunar observer does not see the earth rise or set. If he is on the near



The interval from one new moon to the next is $29\frac{1}{2}$ days, but the time it takes the moon to make one complete orbit of the earth is only $27\frac{1}{3}$ days. The moon falls behind (bottom illustration) because by the time the moon has completed one orbit, the earth has moved along its orbit. It then takes two more days before the moon is between the earth and the sun and we again have "new moon."

side of the moon, the earth is always visible, and if he is on the far side of the moon, he cannot see the earth at all.

When the observer on the moon is at lunar midnight, he sees a "full earth," completely illuminated, unless the moon is directly in line between sun and earth. In that case he observes an eclipse of the earth by the moon at the same time when earth observers see a solar eclipse. When the observer is at lunar midday, and the sun is directly behind the earth, he observes "new earth." If the sun is completely eclipsed by the earth, the earth would be seen surrounded by a bright halo caused by the sunlight scattered in the earth's atmosphere. Conditions for such an eclipse are very similar to those for a solar eclipse seen from the earth, but the track of totality on the moon would be two and one-half times as wide. Because the apparent size of the earth is so much larger than that of the sun when seen from the moon, a lunar observer can never witness the beautiful solar corona or Bailey's beads, which are such striking

features of a total solar eclipse for us. Even if the apparent size of the earth were small enough to cover the solar disk as does the moon, the earth's atmosphere would inhibit the formation of either the "diamond ring" or Bailey's beads, since these also depend largely on the fact that the moon has no atmosphere. Bailey's beads are blobs of light surrounding the moon as it passes across the sun's face, during a solar eclipse witnessed from the earth. They are caused by sunlight shining through deep depressions and between mountain peaks along the edge of the lunar disk. The "diamond ring" is a great halo of sunlight that momentarily appears when the moon covers the solar disk and again when it begins to pass off the disk, during an eclipse of the sun.

Like the moon seen from the earth, the earth seen from the moon passes from "new earth" through crescent, gibbous, and full phases, and back again to "new earth." These changes take the same time as the corresponding changes in the moon seen from the earth, and run through a full cycle in one lunation.

THE MOON AS AN OBSERVING STATION

From the point of view of the astronomer, the moon would be an ideal site for a telescope. The useful size of telescopes on earth is limited on the one hand by our atmosphere, which blurs optical images and limits the kinds of light waves reaching the earth. On the other hand, the engineering problems presented by the construction and operation of large and heavy instruments in the gravitational field of the earth are formidable. In the absence of an atmosphere—as on the moon—the detail that a telescope can record is limited only by the optical properties of the instrument. The smaller value of gravity on the moon would make it possible to operate larger instruments than are feasible on the earth. A lunar observing period for optical instruments would be one-half month long, followed by a similar period in sunlight. Weather would never interfere with observing schedules, but instruments would have to be well protected to withstand the lunar temperature extremes.

MARS

Next to the earth in the order of increasing distance from the sun lies Mars, the red planet, named after the Roman god of war. Mars is a good deal smaller than the earth. Its mean diameter—6,780 kilometers—is a little more than half that of the earth. Its surface area is a little more than one quarter of the earth's surface area. Its volume is only about one seventh that of our planet. The mass of Mars, compared to that of the earth, is 0.11. Its density, compared to the earth's, is 0.70.

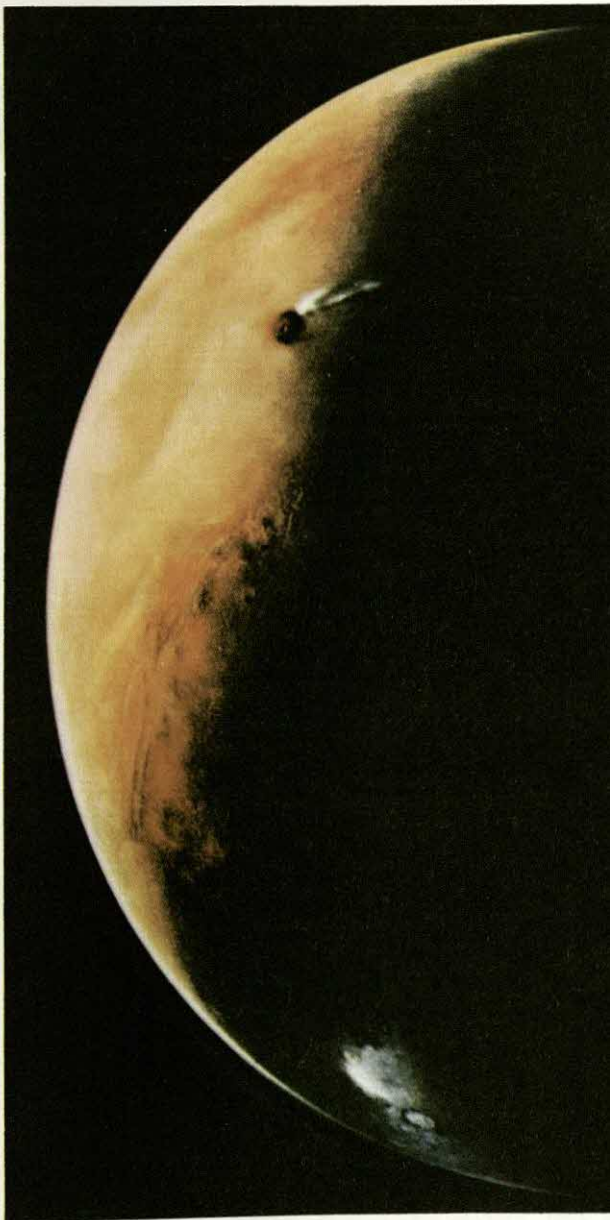
The amount of light and heat that Mars receives from the sun is, surface for surface, less than one-half that received by the earth. As one would expect, Mars is quite a bit colder than our planet. It has been estimated that the temperatures on Mars range from extremes cold enough to freeze carbon-dioxide gas into "dry ice" to those hot enough to melt ordinary ice.

The mean distance of Mars from the sun is about 228,000,000 kilometers, as compared with the earth's mean distance from the sun of approximately 149,600,000 kilometers. At perihelion—the nearest point to the sun—Mars lies about 203,000,000 kilometers away from it. At aphelion—the farthest point from the sun—the distance is about 250,000,000 kilometers. At perihelion, therefore, Mars is some 47,000,000 kilometers nearer to the sun than at aphelion. That means that its elliptical orbit is quite eccentric. The more an ellipse departs from the shape of a circle, the more eccentric it is said to be.

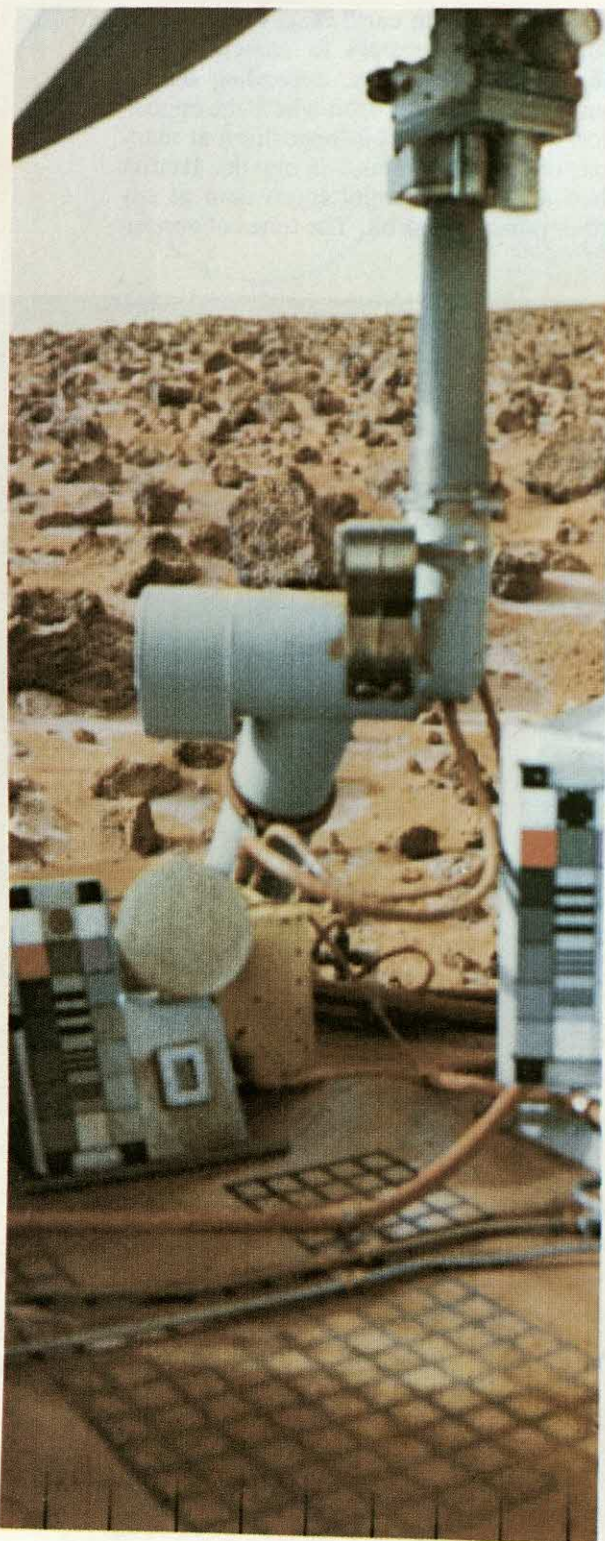
The distance between Mars and the earth varies within very wide limits. When Mars is in conjunction—that is, on the other side of the sun from the earth—its distance from us averages 377,000,000 kilometers. When it is in opposition—that is, on the other side of the earth from the sun—its

distance from the earth ranges from about 56,000,000 kilometers to something like 98,000,000 kilometers, depending on the point in the Martian orbit where the opposition occurs. Mars is in opposition at intervals of 780 days—about 26 months. Its disk then appears far larger to us than at any other point in its orbit. The times of opposi-

NASA



The photo at right was taken by Orbiter II as it approached the dawn side of Mars. The light and dark areas are called continents and seas, although we now know that the planet is dry except for the frozen water of its polar ice caps.



NASA

tion, therefore, give particularly favorable opportunities for carefully examining the surface of this planet.

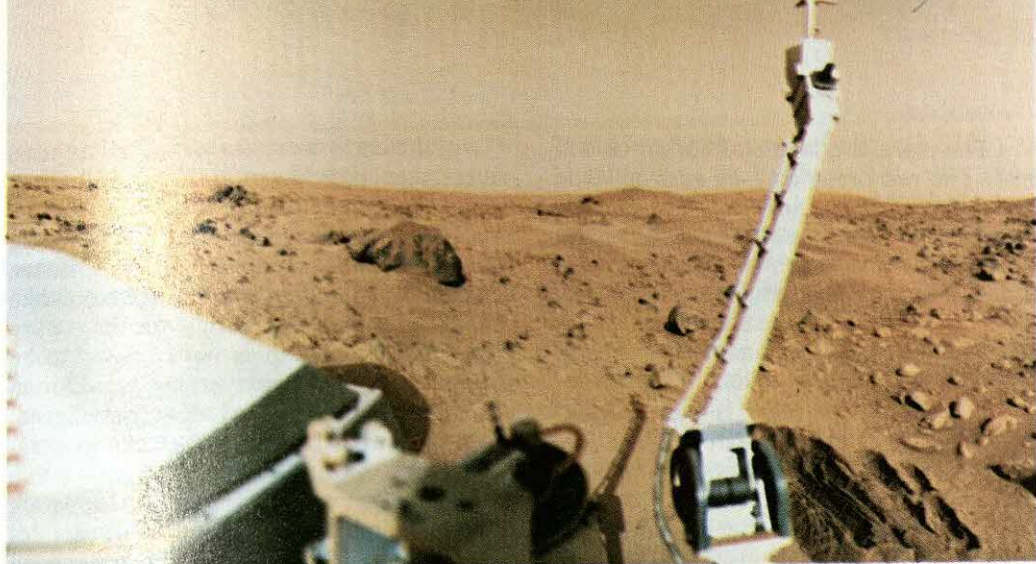
The apparent diameter of Mars varies from about three and a half seconds of arc at conjunction to a little less than twenty-five seconds of arc at opposition. When nearest to the earth, Mars has three times the brilliancy of Sirius, the brightest star in the heavens (with the exception of the sun). When farthest from the earth, its brightness is reduced to that of a star of the second magnitude.

Mars completes its orbit around the sun in 687 of our days (one year and ten and a half months), traveling along its path at the mean rate of 24 kilometers a second. The orbit is inclined to the ecliptic—the plane of the earth's orbit—by less than 2° . The planet rotates on its axis. The clear markings on its surface have made it possible to determine the speed of that rotation with great accuracy. The Martian day is 24 hours, 37 minutes, and 23 seconds long.

Like the earth, Mars is somewhat compressed at the poles. Its equator is inclined by something like 25 degrees to the plane of the planet's orbit. The inclination of the earth to its orbital plane is only a little less— $23\frac{1}{2}$ degrees instead of 25. This tilt accounts for seasonal changes on the earth. In the case of Mars, too, the corresponding tilt brings about changes of season. Since it takes Mars almost twice as long as the earth to complete an orbit around the sun, each of the Martian seasons is nearly six months in length—that is, almost twice as long as the corresponding season of the earth.

The albedo of Mars is 0.15; that is, it reflects 15 per cent of the light that it receives from the sun. Area for area, the disk of Mars is a better reflector than that of the moon, twice as good as that of Mercury, but far inferior to that of Venus. Therefore, it is much less bright than Venus. Like the

This photo shows some of the apparatus of the Viking II lander on the rocky surface of Mars, the red planet. Here we can see patches of late winter frost on the ground. The Martian winter, like its summer, lasts about six months.



NASA

Viking landers probe the Martian landscape. Both Viking I and II landed on relatively flat, rocky plains and took soil and rock samples, in addition to analyzing the atmosphere. Their studies showed the atmosphere to be comprised largely of carbon dioxide and hydrogen gas, and the soil to contain large amounts of oxygen.

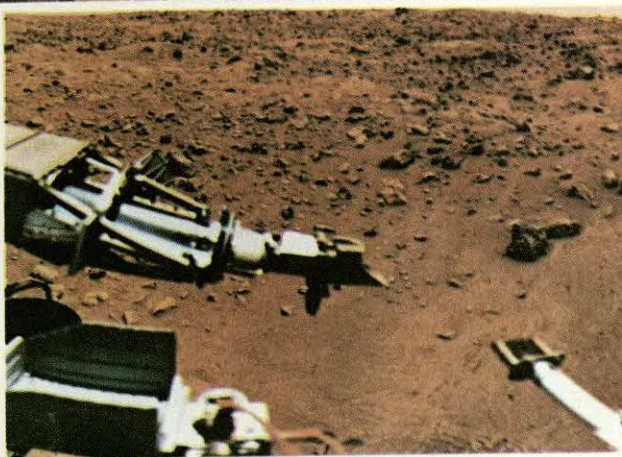
rest of the planets outside the orbit of the earth, Mars does not have phases.

FIRST IMPRESSIONS

Except for the earth, the moon, and perhaps Mercury, Mars is the only solid body in the solar system whose surface we can see. This planet has been under telescopic observation for over three centuries. In the telescope, Mars shows as a small disk with red, dark, and white markings. The red areas, which cover nearly three fourths of the surface, are called *continentes* (Latin for "mainlands"). The dark regions are the *maria* ("seas" in Latin). Around each geographic pole is a white polar cap.

For many years astronomers thought Mars to be much like the earth, with oxygen, water, and polar ice. Somewhat later they realized that Mars is dry, with no large or visible bodies of water at all.

During the 1800's some astronomers claimed they saw many fine lines crisscrossing most of the Martian surface. These lines were named *canali*—Italian for "channels," or "waterways." But to other astronomers the word soon came to mean



NASA

"canals." Astronomers said the red tracts were vast deserts, where almost nothing grew. The soil was supposed to be rust, a compound of iron, oxygen, and water.

During the twentieth century, astronomers gradually realized that this picture of Mars was mostly wrong. Improved telescopes and cameras showed that no canals exist. Instead, there are huge cracks and other natural features arranged in lines.

A CLEARER PICTURE

Detailed studies of the surface of Mars began with the landing of the U.S. space probes Viking 1 and 2 on the planet. Viking 1 landed on July 20, 1976. Viking 2 followed, landing 7,400 kilometers from the Viking 1 site on September 3. These probes and the earlier Mariner probes have given us a much clearer picture of Mars.

The maria are layers of dark rock and dust. The red areas may be rust or some completely different minerals. The familiar markings of Mars seen in telescopes and photographs from earth have little basis in actual Martian topography. Radar and space probe studies have shown, for example, that some maria extend across stretches of varying height, or relief.

The surface of Mars is more rugged than the earth's. There is hardly a land formation on the earth that is not also present on Mars. Mars has high mountains and plateaus, huge volcanoes, craters many kilometers across, broad plains, valleys, steep cliffs, jagged ridges, canyons deeper than our Grand Canyon, sand dunes, long scratches, and cracks, or faults, extending for great distances. Some of the mountain peaks are very high. The volcanic Olympus Mons is, for example, taller than any surface feature on earth. Mars does not, however, have any mountain ranges like those found on earth.

As the sun rises, bright clouds of water ice are seen resting above the canyons of this high plateau region called the "Labyrinth of the Night."

NASA



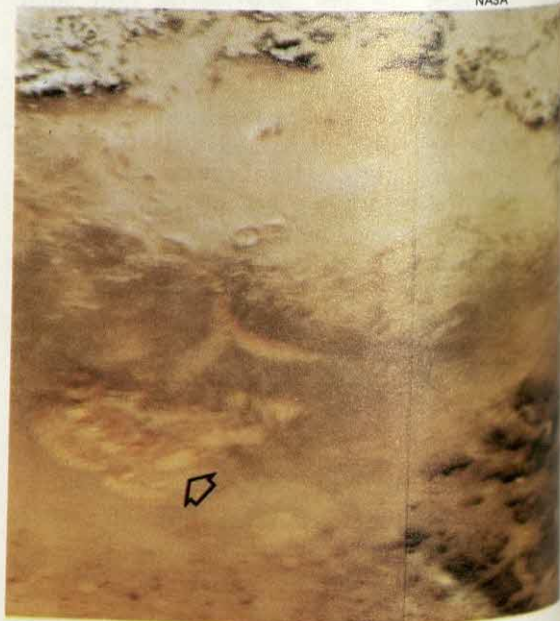
If the earth were drained of all its seas, rivers, and lakes, it would then be at least as rugged as Mars. The empty ocean basins would form deep depressions, with long rifts and jagged ridges along their bottoms. Certain Martian cracks and ridges look like those on the floors of earthly oceans.

Mars' polar cap is now known to be composed of frozen water—not carbon dioxide ice as was previously believed. Some scientists believe that the entire planet may have a layer of permafrost—a thin layer of frozen water—under its apparent surface of dust and rock. Soil samples taken by the Viking probes yielded a good deal of water upon heating.

Was Mars once covered by seas or other bodies of water? Some scientists think so. They point to Martian valleys, gullies, and deposits that seem as if they could only have been caused by streams. Some plateaus, also, are marked by long grooves and scratches resembling those made by glaciers on earth.

A Martian dust storm (arrow), more than 300 kilometers across is viewed moving eastward inside the Argyre Basin, in the southern hemisphere.

NASA



According to one theory attempting to explain these features, Mars' axis of rotation shifts slowly with respect to the sun. As a result, once every few tens of thousands of years, Mars' polar caps become warm enough to melt. They thus release water and moving ice sufficient to erode the planet's surface.

Early Mariner pictures of Mars did not show many of the features discussed above. They did reveal numerous craters, suggesting a dead, pitted surface, like our moon's. Many of the craters dotting the face of Mars are volcanic. Others were dug by falling meteorites.

Changes in the shape and color of Martian surface markings have been seen through telescopes and in Mariner and Viking pictures. Many are undoubtedly caused by the action of meteorites, wind, ice, or volcanoes. Other changes, however, cannot yet be explained in these ways. There may be forces at work on Mars that have no counterparts on earth or on the moon. Some scientists suggest that at least some of the changes may be due to primitive plants growing in some damp, warm spots on the planet.

Instrumental studies show that Martian atmospheric dust and surface rocks

contain a wide range of minerals. They resemble earth rocks chemically, thus showing that Mars, like the earth, has had a long, complex development. Like the earth also, Mars probably has a dense core at its center, which produces a magnetic field.

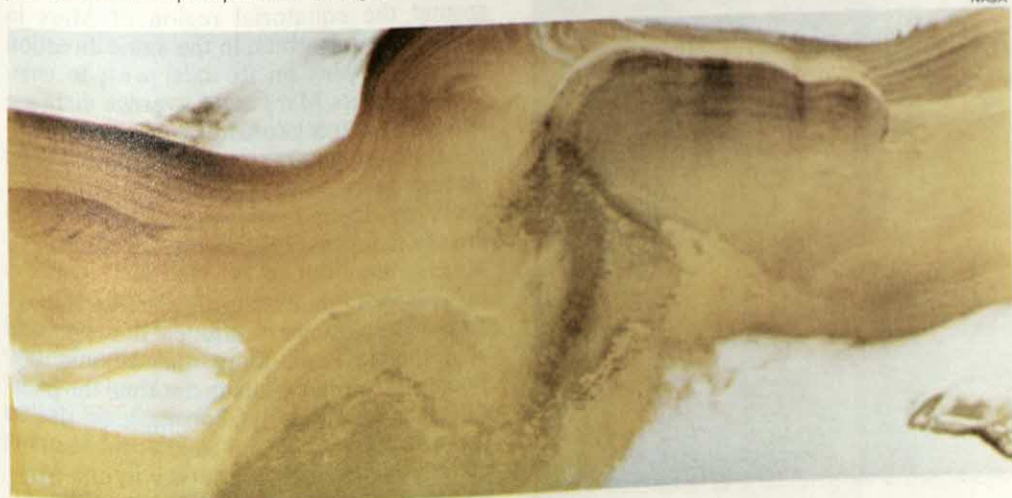
The Viking probes have revealed many interesting facts about the chemistry of the soil on Mars. Large amounts of oxygen are present in the soil, and the soil has a higher-than-expected degree of radioactivity. Some preliminary tests showed that organic matter was present in about one part per million of soil. Later tests, however, seemed to contradict this finding.

MARS' ATMOSPHERE

That Mars is surrounded by air has been known for a long time. What could only be explained as atmospheric hazes, clouds, and dust storms have often been observed there. Mars' blanket of air is too thin, however, to protect much of its surface from the cold and radiation of space.

Air pressure at Mars' surface equals that at many kilometers above the earth's tallest mountains. The top of the atmosphere of Mars is much lower than that of earth's and often lower than the tops of numerous Martian highlands. The iono-

Summer near Mars' north pole: the seasonal carbon-dioxide polar cap clears to reveal water ice and layered terrain beneath. The regularity of the layers suggests a relationship to periodic changes in the planet's orbit.



NASA

sphere—layers of electrically charged gases high in the Martian atmosphere—is weak, so that dangerous radiation from the sun and from space easily reaches the ground.

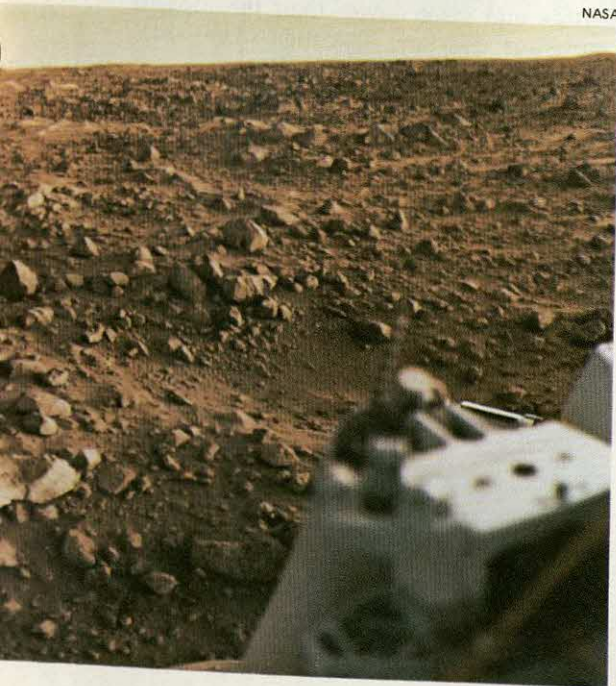
Mars is a cold world. Along the ground at night, especially in the higher latitudes and polar regions in winter, temperatures drop many degrees below the freezing point of water. However, even by earthly standards, Mars at times may be surprisingly mild. The upper Martian atmosphere, for example, is decidedly warmer than ordinary room temperature. At noon, particularly in summer in the lower latitudes and at the equator, surface temperatures may rise far above freezing. Average air temperatures also change over several years.

Chemically, the Martian atmosphere is very different from earthly air. There are only faint traces of oxygen, nitrogen, and water vapor in the atmosphere. In fact, carbon dioxide is the chief gas. A great mass of hydrogen also surrounds the planet.

Thin clouds often form in the Martian

Viking Lander I took this picture of the Martian surface 15 minutes before sunset. A Martian day is just slightly longer than a day here on earth.

NASA



atmosphere. They may be composed of dust, dry-snow crystals, and frozen water. Clouds rise daily over high elevations. They may come from volcanoes or may simply be condensed atmospheric vapors.

The most spectacular and puzzling features of the Martian atmosphere are the gigantic dust storms that periodically sweep the entire planet. The Mariner 9 probe arrived during the height of such a tempest. Winds averaging about 280 kilometers per hour whip up enormous clouds of dust from the desertlike surface of Mars. A storm of this kind may last for weeks or months. During this time, the planet is one vast dust bowl. By comparison, the worst Sahara sandstorm on earth is only a breeze. The Martian surface must be deeply eroded by windblown dust.

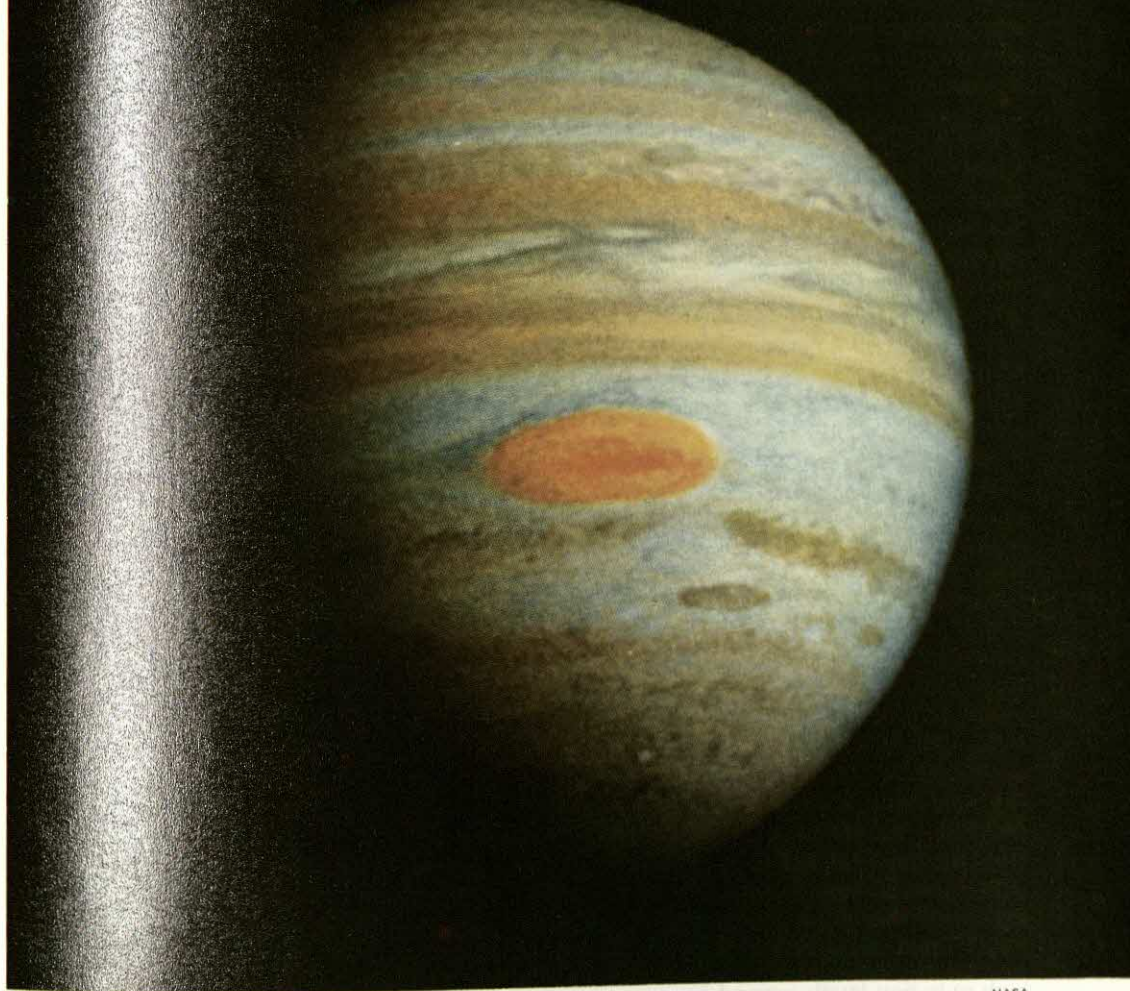
MOONS OF MARS

Mars has two satellites, or moons. They were discovered in 1877 by the American astronomer Asaph Hall. He named them Deimos and Phobos ("Terror" and "Fear") after the two mythical sons and attendants of the ancient god Mars.

Both moons are irregular in shape and very small. Deimos, the outer one, measures about 9 to 11 kilometers. Phobos, the inner one, about 16 to 22 kilometers. Pictures show them to be dark rocky bodies, pitted with craters. They are like asteroids.

Deimos and Phobos both revolve around the equatorial region of Mars in nearly circular orbits, in the same direction that Mars spins on its axis: west to east. Deimos orbits Mars at an average distance of about 19,300 kilometers above the planet's surface. It takes 30 hours and 18 minutes to go once around Mars completely. On Mars, Deimos would be seen to cross the sky from east to west once in about $2\frac{1}{2}$ Martian days.

Phobos is only some 6,000 kilometers above the Martian surface. It orbits the planet once in 7 hours and 39 minutes. Because it circles Mars faster than the planet rotates, Phobos is seen there to rise in the west and set in the east. It makes two complete crossings of the sky in one Martian day.



NASA

The Great Red Spot, the vortex of a storm, is Jupiter's foremost feature. Pioneer 11 took this photo 1,100,000 kilometers from the surface during its flyby.

JUPITER

It is the largest planet in our solar system, bigger than many stars. Its volume is 1,300 times that of the earth. It has a diameter of 142,860 kilometers. In comparison, the earth's diameter is less than 13,000 kilometers — scarcely $\frac{1}{11}$ of Jupiter's.

Despite its huge bulk, Jupiter spins on its axis much faster than our planet, making one complete turn in slightly under 10 hours. This fact explains why Jupiter is noticeably flattened at its north and south poles and bulges around its equator.

For all its enormous size, however, Jupiter is only 318 times more massive than

the earth. Because it has more mass, Jupiter's gravitational pull is 2.65 times that of the earth. Thus if you weigh 50 kilograms, on Jupiter you would weigh 2.65 times as much — more than 130 kilograms.

Scientists have also calculated that Jupiter has only $\frac{1}{4}$ the density of the earth. In fact, its density is only about $1\frac{1}{3}$ times that of water. The reason for this is that the giant planet consists mostly of gases — hydrogen, helium, methane, and ammonia.

Other facts about Jupiter are of interest. It travels around the sun, between the orbits of Mars and Saturn, at an average



NASA

This photograph of the Southern Hemisphere of Jupiter was obtained by Voyager 2. Seen in front of the turbulent clouds near the Great Red Spot is Io, the innermost of Jupiter's larger moons. Io is about the size of Earth's moon.

distance from the sun of nearly 779 million kilometers. Because of this vast distance, Jupiter takes nearly 12 years to go once completely around the sun. At its closest approach to us, Jupiter is about 591 million kilometers away. It is 966 million kilometers away when farthest from the earth.

Pioneer 10 indicated that Jupiter's magnetic field extends out several million kilometers. This magnetic field traps charged particles that make up the "solar wind" that flows out continuously from the sun. These particles result in extremely high radiation levels around Jupiter.

Voyager 1 showed Jupiter to be surrounded by a ring. Some 48,000 kilometers above the Jovian cloud tops, the ring is about 30 kilometers thick and 9,000 kilometers wide. It is too faint to be seen from the earth.

JUPITER'S APPEARANCE

Jupiter often looks bright to the unaided eye. Among the planets, only Venus and Mars may be brighter. Despite its great distances from us, Jupiter is easily visible. It is very large and reflects more than 70 percent of the sunlight falling on it.

Even a small telescope shows the planet as a disk, sometimes surrounded by four spots of light. These are the four "Galilean satellites," the largest of Jupiter's moons. They were discovered in 1610 by the Italian scientist Galileo Galilei, who was the first to observe Jupiter with a telescope.

In large telescopes and in the Pioneer photographs, Jupiter is a spectacular sight. Its wide disk is crossed by many bands of color—pastel shades of blue, brown, pink, red, orange, and yellow. These bands, or belts as astronomers call them, are parallel to Jupiter's equator.

Jupiter's belts, despite changes in their colors and other features taking place from time to time, are mostly permanent. Jupiter's face is also marked by shifting patches and spots, the most obvious of which is the famous Great Red Spot.

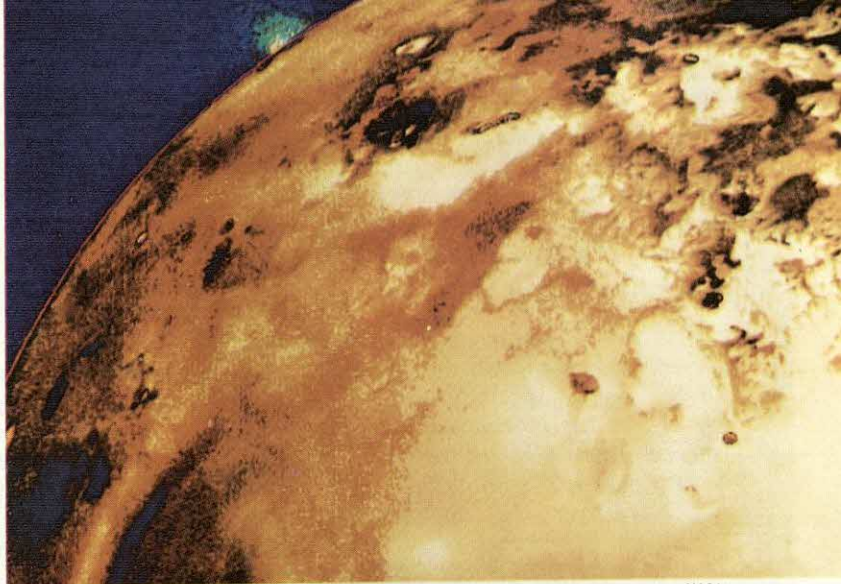
PHYSICAL CONDITIONS

The "surface" of Jupiter seen in telescopes and photographs is not actually a solid surface or crust. In fact, we are not sure whether Jupiter has any solid part at all. It may be a gigantic ball of substances that are usually known as gases on the earth. But there may be at least a small, rock-hard or partly solid core at the center of the planet.

The multicolored belts at Jupiter's "surface" are layers of dense atmospheric clouds, composed of liquid drops and frozen particles. Just how deep the atmosphere and its clouds extend is not certain, but the depth is probably about 1,000 kilometers. Toward the interior of Jupiter, the clouds become heavier, and the gases assume a liquid form.

About 24,000 kilometers from the surface, the liquified gases become truly solid,

A volcanic explosion on Io as photographed by Voyager 1. Solid material had been thrown about 160 kilometers high at an ejection velocity of about 2,000 kilometers per hour. The pink dots are not related to the explosion.



NASA

because of the enormous pressures, thousands of times those at the surface of the earth. The conditions just described would explain the low average density of Jupiter.

The idea that Jupiter's visible face is fluid agrees with the way the planet rotates. Depending on latitude and feature, Jupiter makes a complete turn on its axis in from 9 hours 50 minutes to almost 9 hours 56 minutes. The equatorial region spins fastest, the polar regions slowest. This is the way a mass of gas or liquid would rotate. If there is a solid core, one would expect the planet to rotate more uniformly.

Because of Jupiter's rapid spin, great distance from the sun, and deep clouds, the planet's weather must be strange and different indeed from the weather on the earth. Yet certain cloud formations in the earth's atmosphere resemble the belts of Jupiter. And the planet has, like the earth, great cyclonic storms, or regions of low atmospheric pressures, as well as their opposites: anticyclonic flows with high pressures.

Jupiter radiates two or three times as much energy as it receives from the sun. This means that the planet has an internal source of energy—probably the energy left over from the time when it was first formed. Some astronomers think of Jupiter as a star that simply didn't "make it" because it was too small. That is, its mass was too small to produce the internal pressures and temperatures needed to set off the nuclear reactions that take place inside a star.

Jupiter also radiates energy in the form of radio waves. Some of them originate in atomic particles that come from the sun and are trapped in a belt around Jupiter by the planet's magnetic field. Scientists use radio telescopes to study the waves.

CHEMICAL COMPOSITION

If you landed on Jupiter and tried to breath the air, you would soon die. The substances in Jupiter's atmosphere would poison or suffocate most living things from the earth.

These dangerous substances are the elements hydrogen and helium and the hydrogen-carrying compounds methane, ammonia, and possibly hydrogen sulfide; on earth they are normally gases. Jupiter lacks the chief ingredients of the earth's atmosphere: free, molecular oxygen, nitrogen, and carbon dioxide. But the planet probably has water.

If you somehow could survive for a while in Jupiter's atmosphere, you would smell extremely foul odors. They come from the ammonia and the hydrogen sulfide. Ammonia is a pungent compound of hydrogen and nitrogen. On earth it is produced by the decay of certain proteins. Hydrogen sulfide, another protein-decay product, contains sulfur and smells like rotten eggs.

The other chief constituents of Jupiter's atmosphere—hydrogen, helium, methane, and water—are odorless. Methane is a

chemical compound of hydrogen and carbon. It is formed on the earth by the decay of organic matter in the absence of atmospheric oxygen. It is also called marsh gas, because it frequently arises in marshlands.

THE GREAT RED SPOT

Almost from the time astronomers first observed Jupiter with the telescope, over three centuries ago, they have been puzzled by one of the strangest features ever observed on any body of the solar system. It is the Great Red Spot.

Located at 20° latitude in Jupiter's Southern Hemisphere, the Great Red Spot has an oval shape and measures about 50,000 kilometers at its widest. Its size varies. Also, it changes in color from a bright brick red to an almost invisible pink and then back, over periods of many years.

Although the spot maintains more or

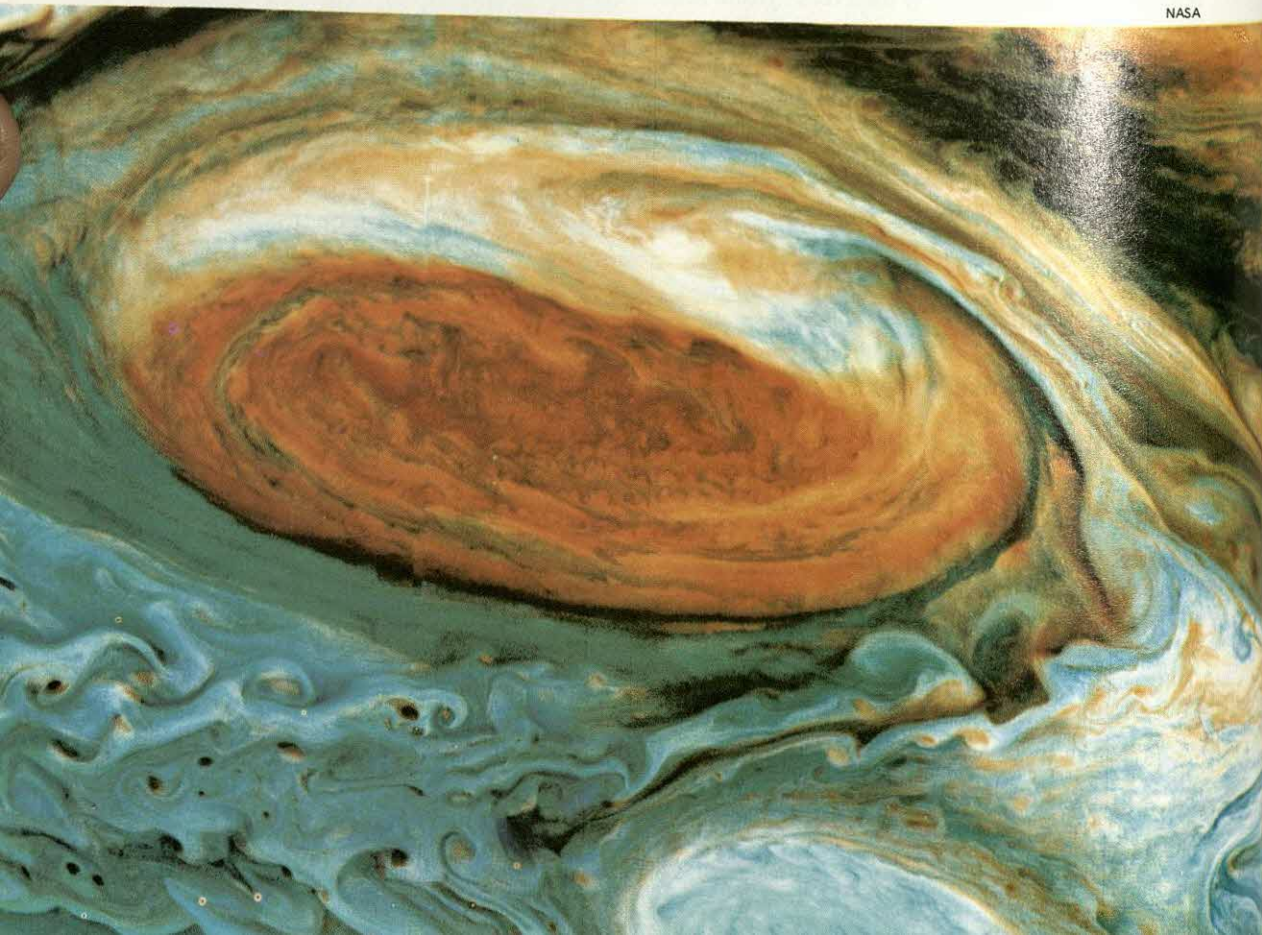
less the same latitude, it may change its longitudinal position markedly. Also, the Great Red Spot has a speed of rotation around Jupiter's axis averaging a few kilometers per hour slower than those of the areas close to the spot. It may move slightly faster or slower from time to time.

It is now the general opinion of space scientists that the Great Red Spot is a violent storm raging in the upper atmosphere of Jupiter. We are looking down upon a hurricane-type event so enormous that it could easily swallow up the entire earth. The storm has raged for many centuries and gives no signs of coming to an end for many more.

The close-up photographs taken by Pioneer 10 and 11 show the "pinwheel" structure of the Great Red Spot. It is spinning so rapidly that it is rigid enough to displace the clouds of the south tropical zone.

This Voyager 1 view of the Great Red Spot is shown in enhanced color, which emphasizes red and blue over green. At the bottom is one of three oval cloud systems that formed nearly 40 years ago.

NASA



This color-added photo of Callisto, the outermost of Jupiter's inner moons, was taken by Voyager 2. The icy surface of Callisto is the most densely cratered of Jupiter's moons. The bright areas are probably material thrown out by relatively recent impacts with meteorites.



NASA

Temperature measurements indicate that the storm rises some eight kilometers above the surrounding cloud deck.

The photographs also revealed the presence of other, smaller storm centers, such as the Little Red Spot in the Northern Hemisphere of Jupiter. In fact, the face of Jupiter as a whole is a scene of much weather activity. The grayish white zones are warm, upward-rising weather "cells" that carry heat from the interior of the planet to the surface. The lower, red-brown areas are downward-sinking cells.

THE INNER MOONS OF JUPITER

Jupiter has 16 moons—four of which are several thousand kilometers in diameter. Some of the small, outer moons may be former asteroids, captured by the planet's vast gravitational field. The densities of the moons decrease outward as the spacing of their orbits increases. The inner moons are rocky, and the outer ones are possibly frozen liquid or gaseous.

Photographs taken from Voyager 1 revealed several new satellites. The newest of these, tentatively called 1979 J3, is only 40 kilometers in diameter and orbits at the edge of the Jovian ring system at a great speed.

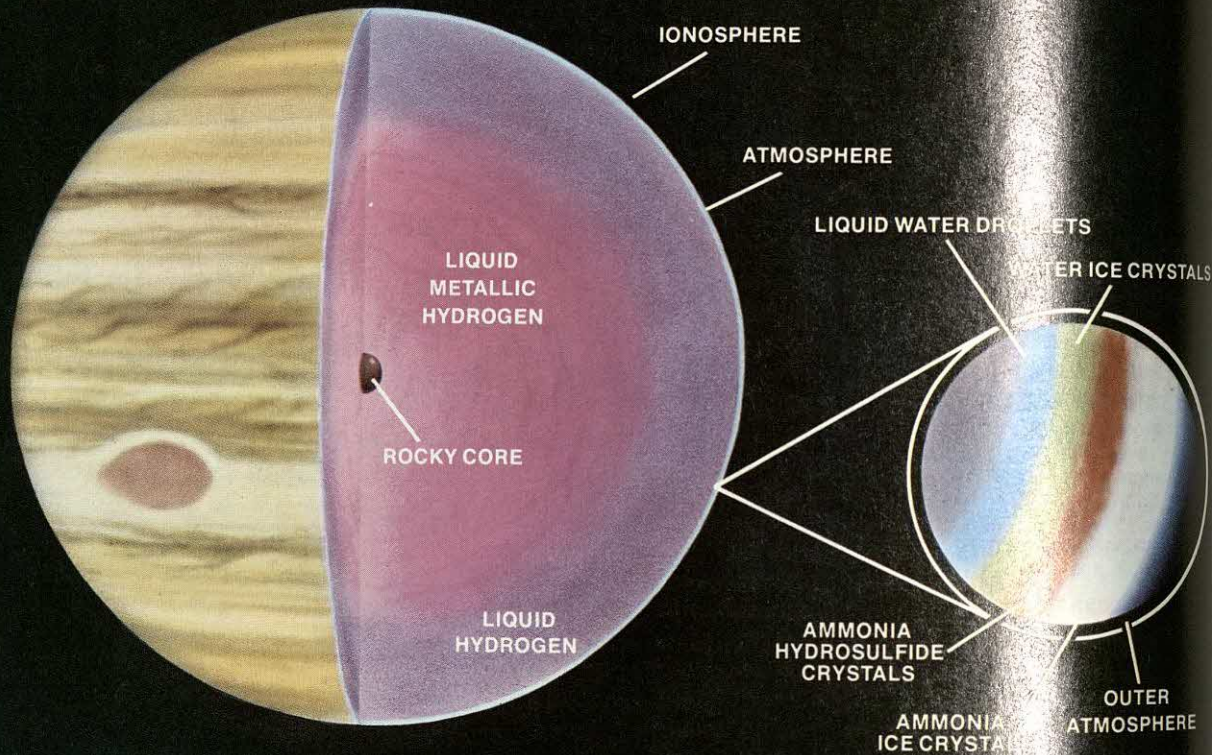
The most interesting moon is Io, which is volcanically active. Voyager 1 passed less than 22,400 kilometers under Io's

south pole, taking numerous photographs of its surface. Io is so active that no permanent features could be distinguished in the Voyager photographs, making it difficult to determine the moon's rotation period. However, the photographs did show several erupting volcanoes spewing dust many kilometers above them. Voyager 1's infrared sensor detected enormous lava lakes whose temperatures are between 78° and 93° Celsius. The surrounding areas are about -162° Celsius.

Inspection showed many surface features also found on earth, such as rounded hills, plateaus, cliffs, depressions, and plains. One scientist compared the surface of Io to Yellowstone National Park, which has been "cooked, steamed, and fumed."

Europa is Jupiter's fifth closest moon. Voyager 2 photographs of it show relatively few impact craters. But it is marked by long linear features thousands of kilometers long and tens of kilometers wide. These features appear dark on its light, icy surface.

Ganymede is Jupiter's largest moon. It is 5,270 kilometers in diameter. Our moon's diameter is 3,480 kilometers by comparison. Voyager 2 came within 60,000 kilometers of Ganymede, photographing its earthlike surface. The surface shows numerous parallel faults (cracks), which appear to be like those on the Pacific and Atlantic



Astronomy, November 1974

This model of Jupiter's interior shows the planet as primarily a ball of liquid hydrogen, perhaps with a small rocky core. The atmosphere, about 1,000 kilometers deep, consists of several layers, as shown in the small closeup.

Ocean basins. There are also areas that have been pounded by meteorites, which have left circular impact craters.

Callisto, which is almost as large as Ganymede, shows an icy surface that has preserved the heavy meteorite bombardment of the early period of the solar system. Nothing appears to have happened to Callisto since the final formation of this moon. The craters appear to be rounded. There appears to be no kilometer-high crater walls and towering peaks as on earth's moon.

LIFE ON JUPITER?

For many years, astronomers were convinced that nothing could live on Jupiter. They pointed to the unearthly conditions: the brutal climate, the lack of oxygen, and the presence of poisonous or suffocating chemicals. Today the attitude of scientists is completely different. These very

conditions are now thought to be a breeding ground for substances that could develop into living matter.

As we have said, methane, ammonia, and hydrogen sulfide, which are abundant on Jupiter, often arise on earth from decay of organic matter. Does this mean that Jupiter is filled with decaying organic matter? Not at all.

Suppose we could reverse the process of decay. What if we could take ammonia, methane, hydrogen sulfide, water, and other chemicals, combine them in different ways, and get proteins and other biological compounds? According to many scientists, this may be happening on Jupiter. Furthermore, this process may have begun to take place on the earth 4,000,000,000 to 5,000,000,000 years ago. Studies have indicated that, although life flourishes on earth today, it could never have arisen from ordinary matter under our present conditions.

SATURN

For beauty and interest alike, there are few objects in the starry heavens to compare with Saturn. This magnificent planet, with the system of rings that encircles it, provided an unforgettable spectacle when viewed by the U.S. Voyagers 1 and 2 spacecraft in 1980 and 1981. The Saturnian system includes not only the planet and its rings—probably thousands of rings, ringlets, and subringlets—but also as many as 21 satellites, or moons.

Until the days of the great 18th-

century English astronomer Sir William Herschel, Saturn was considered to be the outermost of the planets. To be sure, Uranus, a planet far outside the orbit of Saturn, is clearly though faintly visible to the naked eye. But it was not revealed as a planet until 1781, when Sir William discovered that what was once thought to be a star had a definite disk. Today we know that there are at least two other planets outside the orbit of Saturn—Neptune and Pluto.

To the ancients Saturn appeared to be

The rings of the magnificent planet Saturn are bright and the features of the planet's northern hemisphere are well-defined in this Voyager 2 photo taken on July 12, 1981 from a distance of 43 million kilometers.

NASA



the most insignificant of the heavenly bodies that were supposed to circle the earth (the sun, the moon, Mercury, Venus, Mars, Jupiter, and Saturn), as distinguished from the fixed stars. The glorious rings that surround the planet were invisible before the invention of the telescope in the first decade of the seventeenth century. Otherwise this magnificent crown might have saved Saturn from the sinister reputation that it once bore. Ancient astrologers maintained that it had a sinister influence upon people.

It is believed that Saturn acquired its bad reputation because of the slowness with which it moved against the background of the fixed stars and also because it casts a comparatively dim light. To be sure, it appears to the naked eye as a very bright star, but it is far less bright than Venus, Mars, and Jupiter, and only about as bright as little Mercury, which lies nearest to the sun. When the sign of the zodiac where Saturn happened to be appeared near the horizon at the time of a person's birth, that unfortunate person was supposed to be particularly subject to the evil influence of the planet: misfortune would follow ever afterward.

MASSIVE AND REMOTE

Saturn is so far from the center of the solar system that, viewed from its orbit, the sun would appear as a brilliant pinpoint rather than as a disk. Saturn is nearly twice as far from the sun as Jupiter, and receives only one-ninetieth of the heat and light that we receive on earth. So remote is Saturn that from its surface none of the planets within its orbit except Jupiter would be visible. Jupiter would appear as a companion to the sun, sometimes an evening star, and sometimes a morning star, as our sister planet Venus appears to us.

The mean distance of Saturn from the sun is 1,428,000,000 kilometers, or about $9\frac{1}{2}$ times the distance of the earth from the sun. It completes its orbit once every $29\frac{1}{2}$ years. Our own planet, the earth, overtakes it and comes in line between it and the sun once every 378 days—that is, in about a year plus two weeks. Saturn is inclined to

the ecliptic, the apparent path of the sun in the heaven, by $2\frac{1}{2}$ degrees. Its orbit, as it revolves around the sun, is much more eccentric than that of the earth.

The distance of Saturn from the earth varies according to the position of the two planets in their orbits, from 1,197,000,000 kilometers to 1,654,000,000 kilometers—a variation not sufficient to cause any great difference in the planet's brightness.

The globe of Saturn is greatly flattened, so that when the planet is in such a position that the plane of its equator passes through the earth, its profile appears definitely elliptical—a feature to which much of its striking appearance is due. The planet's polar diameter is nearly nine-tenths that of the equatorial diameter—108,000 kilometers and 120,000 kilometers, respectively. These dimensions show the vast size of the planet. Its volume is nearly 750 times that of the earth; its superficial area over eighty times that of our globe.

The density of Saturn is very low, much lower than that of any other planet. In fact it is only about three-quarters that of water. Because of this fact some astronomers hold that Saturn is far from having reached the solid condition. They maintain that it must cool for long ages to come before it will approximate the present condition of the earth. Saturn's flattened shape is believed due to this molten (or vaporous) condition and to its rapid rotation.

Its swift rotation on its axis was first observed by Sir William Herschel in 1794 by means of cloudlike markings visible on the planet. These indicated a rotation period of ten hours and sixteen minutes. In 1876, the American astronomer Asaph Hall noticed a brilliant white spot on Saturn's equator. Using this spot as a point of departure for his calculations, he found that the equatorial period of rotation was ten hours and fourteen minutes, two minutes less than the period determined by Herschel. This spot, which appeared to mark a vast eruption of glowing material from the planet's interior, remained visible for several weeks. During that time many astronomers made a careful study of the planet's daily motion.

In 1981, radio astronomers using data from Voyagers 1 and 2 were finally able to pin down Saturn's rotation rate to 10 hours, 39.9 minutes. The inaccuracies of earlier measurements were due to their being taken from the shifting cloud cover.

The axis of Saturn is inclined to the plane of its orbit by about twenty-seven degrees, giving the planet much the same slant as the earth has in its orbit. The inclination of the earth's axis accounts for its seasons.

Little is known about Saturn's weather and seasons. The seasonal changes on Saturn in the nine and one-half months between the visits of Voyagers 1 and 2 was the equivalent of barely a week on earth. However, weather changes were seen. One large white spot in the Northern Hemisphere had lost its dark border. Huge irregular storm systems had evolved new

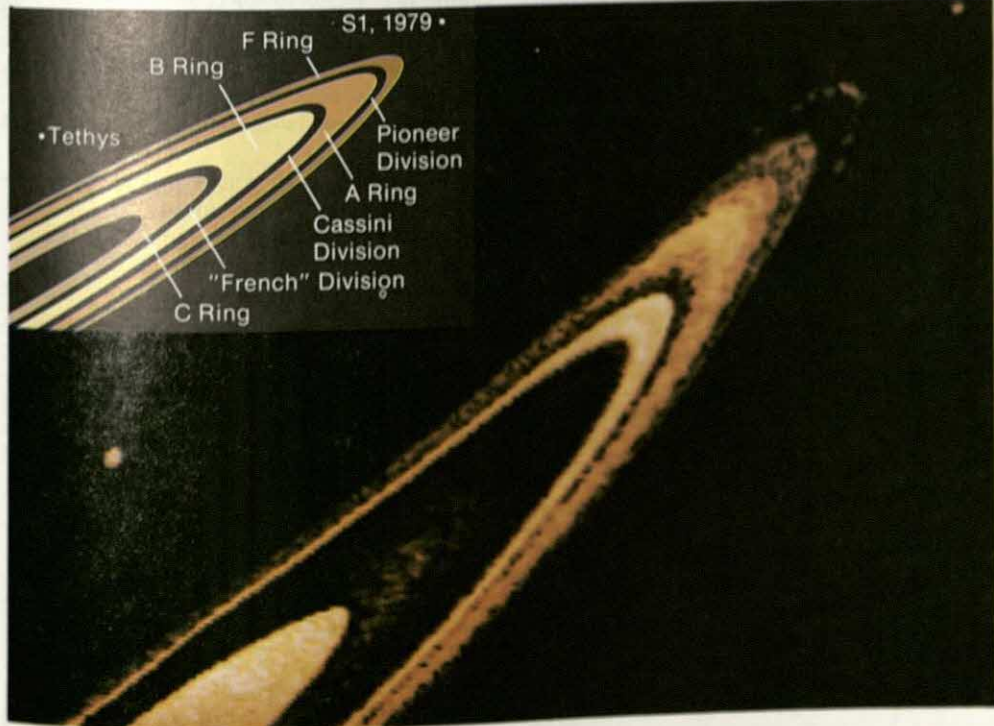
shapes—yet they remained visible, suggesting that on Saturn, as on Jupiter, storms are long-lived.

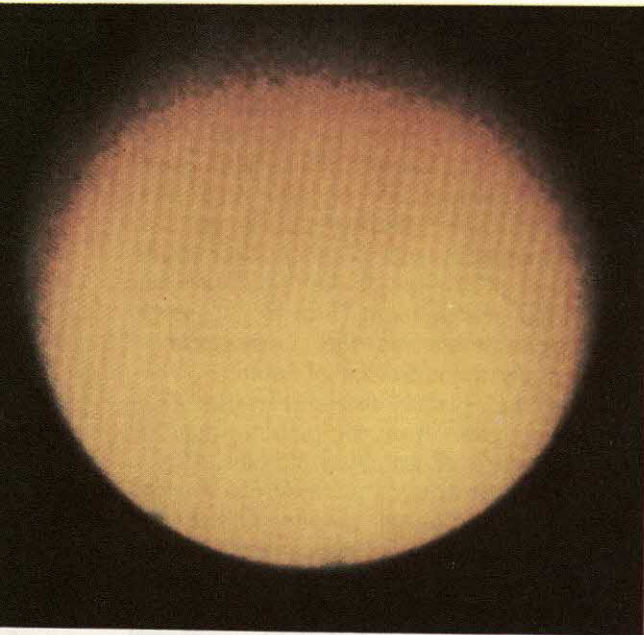
There is a very brilliant and broad white belt around the equator of Saturn. The poles have a cap of dull green. The spectroscope reveals a deep atmosphere. Like that of Jupiter, it contains both ammonia gas and methane (marsh gas). Methane predominates in the atmosphere of Saturn; ammonia, in that of Jupiter.

Voyager probes detected strange ping-pong radio signals from the planet. Sounding like the pings of dolphins playing deep in the ocean, the sounds were detected in the vicinity of the inner moons Tethys and Dione. A doughnut-shaped cloud—a torus—of electrically charged particles surrounding the planet may act as a barrier, containing the radio signals and keeping them from propagating outward. This torus,

Backlit by sunlight, this Pioneer photograph of Saturn's rings shows details never before seen, including the presence of material running through Cassini's division. The art insert identifies the rings and divisions and points out the satellite Tethys.

NASA Ames Research Center





NASA

Voyager 2 obtained this view of cloud-shrouded Titan, Saturn's largest satellite, on August 22, 1981 from a distance of 4.5 million kilometers.



NASA

Iapetus, the outermost of Saturn's large satellites, as seen from Voyager 2 on August 22, 1981 from a distance of 1.1 million kilometers.

also discovered by the Voyager probes, is thought to be similar to the one Voyager had discovered earlier near Jupiter.

SURROUNDED BY PARTICLES

Saturn is surrounded by a vast swarm of small particles. They circle the planet in the form of many—probably thousands of—rings, some quite complex in structure. Saturn's rings were long thought to be unique in the solar system, but in 1977 the planet Uranus was also found to be encircled by a system of rings. Study of Saturn's rings proceeded slowly over the centuries and then took a giant step forward with the surprise findings of the Voyager probes.

It was not until the telescope was discovered, in the first decade of the seventeenth century, that men had the first inkling of the very existence of the rings. When Galileo examined Saturn in 1610 with a telescope that he had made with his own hands, he came to the conclusion that the planet had a triple form. "When I observe Saturn," he wrote to a friend, "the central star appears the largest; two others, one situated to the east, the other to the

west, and on a line that does not coincide with the direction of the zodiac, seem to touch it. They are like two servants who help old Saturn on his way and always remain at his side. With a smaller telescope, the star appears lengthened, and of the shape of an olive."

Continuing to watch this strange performance, month after month, Galileo was amazed to see Saturn's attendants becoming smaller and smaller, until they finally disappeared altogether. He doubted the evidence of his telescope. "What can I say," he wrote, "of so astonishing a metamorphosis? Are the two small stars consumed like sun spots? Have they vanished and flown away? Has Saturn devoured his own children? Or have the glasses cheated me, and many others to whom I have shown these appearances, with illusions?" Discouraged, he abandoned his observations of Saturn.

Others, however, watched the planet whenever it was in view, and they gradually established the fact that Saturn's unusual appendages underwent regular changes. They appeared first as bright, straight lines

stretching outward on either side of the planet's elliptical disk. For the next seven years, these mysterious lines expanded into two luminous crescents attached to the planet like handles to a dish. For the next seven years, the crescents became flattened down until they were again lines projecting from Saturn. Finally they disappeared altogether. For as the planet pursues its vast orbit, slanting always in the same direction, its rings, at opposite points in its orbit, appear in an edgewise position to observers on earth. Between these two edgewise appearances, the north and south faces of the rings are visible in turn, always foreshortened. Each is seen for a period of about 15 years.

In 1655 the Dutch mathematician, physicist, and astronomer Christian Huygens invented an improved method of grinding lenses. As a result he was able to construct a powerful telescope that showed details more clearly than any earlier instrument. Using his new telescope, Huygens observed that the rings of Saturn cast shadows on the planet and were separated from it. From this observation, he deduced the true nature of Galileo's "appendages," which had puzzled astronomers for such a long time.

SATURN'S SEVERAL RINGS

The idea that Saturn had rings was accepted in time. Huygens' fellow astronomers, including the renowned Giovanni Domenico Cassini, began to study the rings more carefully. In 1675 Cassini observed a dark band in what was then believed to be the single ring of Saturn. This band divided the ring into two separate rings. The dividing "band," which was really a gap, has since been labeled "Cassini's division."

A third ring was observed in 1838 by the German astronomer Johann Gottfried Galle. His report of a third ring was ignored by his contemporaries. It was not until the ring was observed and reported simultaneously by W. C. Bond at Harvard and W. R. Dawes in England in 1850 that its existence became an accepted fact. In the third ring, there is apparently much less material that can reflect the sun's light back to us. It has been likened to translucent

crepe paper or gauze. As a matter of fact, it is often called the crepe, or gauze, ring.

In 1837 Johann Franz Encke, the director of the Berlin observatory, saw what he believed to be another division in the outer ring. It was not complete and was not equally distinct at all times. This indistinct and transitory division has been called "Encke's division."

In 1969 the French astronomer Pierre Guerin discovered a fourth faint ring of Saturn, long undetected because it is so close to the bright globe of the planet.

The Voyager probes of 1980 and 1981 revealed more major rings, discovered that the major rings are composed of many ringlets, and determined that the divisions are not empty space at all but rather contain ringlets.

The major rings of Saturn, going from the outermost to the innermost, are designated by the letters E, F, A, B, C, and D.

RING DIMENSIONS

The E, or extended, ring lies outside the other major rings and may be as much as 100 kilometers thick.

The F ring has generated much interest. It is a very narrow, wispy ring, perhaps composed of three to five ringlets that may be braided. Two tiny moons straddle the ring and have been thought to act as "shepherd" moons, confining the ring to its narrow path by a kind of gravitational pinching action. Because Voyager 2 showed the ring undisturbed by the moons' presence, the theory is now questioned.

Ring A is the second brightest ring. It is 16,000 kilometers wide and has an outside diameter of 273,000 kilometers. It may be no more than 100 meters thick. In this ring appears the narrow Encke's division, containing multiple ringlets.

The width of Cassini's division, which lies between Rings A and B, is about 3,500 kilometers. This division is known to contain a series of ringlets, some tightly packed together.

Ring B is the brightest ring. Its width is 26,000 kilometers. Its outside diameter is 235,000 kilometers. Voyager studies revealed that the B ring is composed of

perhaps 300 ringlets, each of which may be made up of some 20 to 50 subringlets.

The biggest surprise involving the B ring, however, was the appearance of "spokes," or "fingers," across the ring. These fingers move with the ring and are thought to be clouds of very fine particles raised above the plane of the ring. Voyager 2 also detected "lightninglike phenomena" from the region of the B ring. Whether these electrical discharges are responsible in some way for the spokes is the subject of much ongoing research.

Ring C is the crepe or gauze ring. It is separated from Ring B by only about 1,600 kilometers. It is 18,500 kilometers wide. Its outside diameter is 196,000 kilometers.

Ring D is the faintest of all. It extends from the surface of Saturn to the inner edge of the C ring.

For many years after the discovery of the rings, astronomers had no idea of their composition. Little by little, as new observations were made, the facts concerning the rings were pieced together. Cassini, who was the first astronomer to discover that there was more than one ring, was also the first to offer a theory concerning the composition of the rings. As early as 1715, he suggested that they were composed of a great many little meteors. This, however, was just a guess.

COMPOSITION OF THE RINGS

In the early 19th century, the French mathematician and astronomer Marquis Pierre Simon de Laplace showed that the rings of Saturn could not possibly be solid or fluid. By applying the laws of motion and force to an imaginary solid ring, he showed that such a ring would be subjected to great stresses and strains from the tremendous force of the revolution around the planet. These stresses and strains would eventually break up the ring. In the case of a fluid ring, by applying the same laws of motion and force, Laplace showed that great waves would be set up in the ring as it revolved around Saturn. As a result the fluid would eventually fly off into space.

More positive evidence was supplied by the celebrated English philosopher and

physicist James Clerk Maxwell. In 1859 he published an essay called "On the Stability of Saturn's Rings," in which he proved mathematically that only if the rings consisted of small satellites could they remain as stable as they had been since they were first discovered.

The intricate nature of the rings was later to be confirmed by the findings of other astronomers. On February 9, 1917, two English astronomers, M. A. Ainslie and J. Knight, reported that a fairly faint star had been observed through the rings, although it had lost most of its brightness. On March 14, 1920, a similar observation was made at the W. Reid observatory in Rondebosch, South Africa. Later the Dutch-American astronomer Gerard P. Kuiper made a study of sunlight reflected from the rings with a device known as the lead sulfide cell. On the basis of his findings, he suggested that the rings may be composed of many small meteoritic ice particles. This suggestion is supported by the latest Pioneer and Voyager observations.

HOW DID THE RINGS ORIGINATE?

In 1850 Edouard Roche proved mathematically that any satellite within a certain distance of a planet around which it rotates would be torn apart by the gravitational forces of the planet. The gravitational force exerted by one body on another is inversely proportional to the distance between them. As they approach one another, the force becomes greater. Roche calculated that the exact distance at which the force of gravity of a planet would be great enough to tear apart its satellite is 2.44 times the radius of the planet. This distance is now known as Roche's limit.

A line drawn from the center of Saturn to the outside edge of the visible ring system would be approximately 2.35 times the radius of the planet. This places the visible ring system inside Roche's limit. All of the known satellites of Saturn whose orbits are sufficiently well determined to establish their distances from the planet are outside Roche's limit. Even the satellites first observed by Pioneer and Voyager probes and believed to be closest to the planet are at

distances at least 2.55 times the radius of Saturn. Roche's calculations thus delimit a specific region within which satellites do not appear.

Here we have a clue to the origin of the rings. According to one theory, a "wandering moon," or satellite, was torn apart by the planet's gravitational forces. The fragments of the former satellite now form the material of the rings. There is another quite similar theory of the origin of the rings. It holds that when the gases in the vicinity of the planet cooled and formed the various solid balls, or satellites, the gases inside Roche's limit were prevented from combining into a satellite by the strong gravitational forces of the planet. Therefore, instead of merging into a single ball, they cooled as small fragments. These fragments revolve around Saturn separately in the form of the rings. The fact that the bona fide satellites of Saturn are all outside Roche's limit seems to support these interesting theories.

SATURN'S SATELLITES

As many as 21 satellites may revolve around Saturn. The largest, Titan, is a planet in its own right, the "Jupiter" of the Saturnian system. The outstanding feature of the smaller moons is that two, even three, of them may share a single orbit.

The two satellites that appear to shepherd the F ring, for example, orbit close to each other and pass each other every few months. S-10 and S-11, and Dione and Dione-B, also appear to be heading toward a similar "dodge-'em" maneuver. Two newer moon-candidates were detected leading and trailing a larger moon, the satellite Tethys.

Tethys is a heavily cratered satellite, as is Mimas, which carries a huge impact scar on its side. Pictures of Tethys reveal a crater one-third the size of the moon, the largest crater seen on any of the Saturnian satellites. Enceladus, on the other hand, appears much smoother. It has empty

If you were standing on the frozen surface of Titan, you might have this spectacular view of Saturn.

Painting by Chesley Bonestell, Griffith Observatory



plains and wrinkled ridges and valleys, suggesting a rather recent history of activity. Some scientists think that a slush may have obliterated impact craters on the moon and made at least part of the surface less than 100,000,000 years old.

Rhea, photographed extensively, is brownish in color and has a crater-saturated surface. The craters are of different sizes, suggesting to some that Rhea may be revealing two stages of impact: one series of impacts from the formation of the Saturnian system and one from debris from the rest of the solar system.

Iapetus and Hyperion were also studied by Voyager the probes. Iapetus was found to be 80 per cent ice with at least half of its surface covered with "a stain of organic material." The moon appears to have a light and dark side, the dark side perhaps resulting from a coating of dust particles. Hyperion held some surprises. Revealed as a "battered hockey puck," this strange misshapen moon, about 190 kilometers by 345 kilometers, may have been smashed into its irregular shape by impact with another body. Its unusual shape and unusual axis orientation are giving scientists much to study.

Titan, the largest of Saturn's moons,

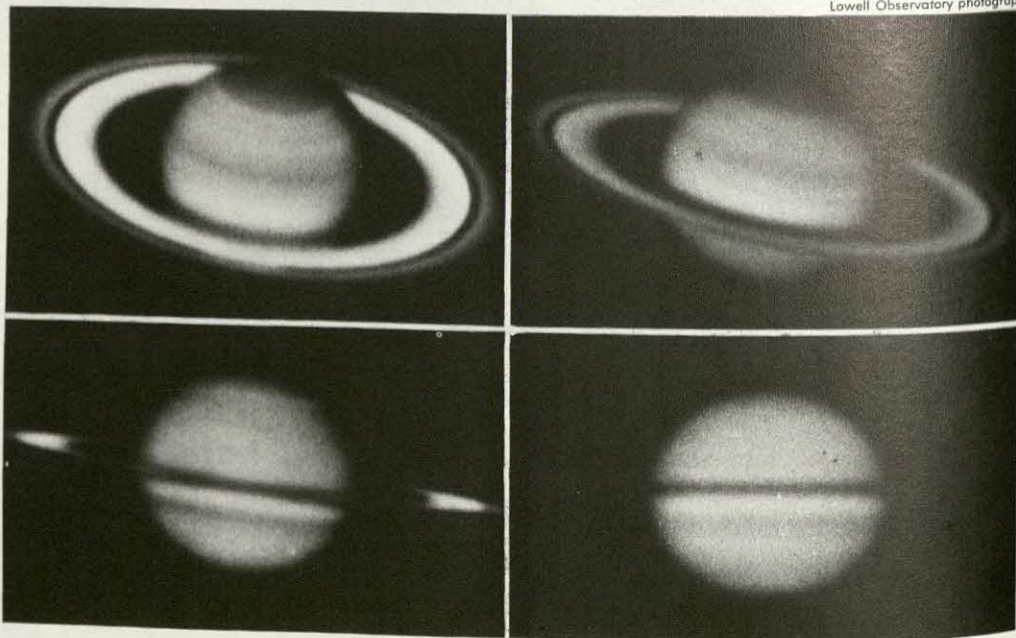
generates the most interest, however. It is nearly twice as large as the earth's moon, or about one-half as massive as the planet Mercury. It is enveloped in a red haze and appears as a nearly featureless reddish-orange disk, slightly darker in the Northern Hemisphere. Its atmosphere was long thought to be mostly methane, but the Voyager probes revealed that it is 99 per cent nitrogen with some methane, acetylene, ethylene, and hydrogen cyanide. The presence of hydrogen cyanide is of particular interest to scientists, for this molecule is believed to be a necessary precursor to the chemical evolution of life. In many ways Titan mimics the conditions present on early earth and scientists hope that the study of Titan may reveal much about how life began and evolved on earth.

YEARS OF STUDY

The data gathered by the two Voyager probes have provided a tremendous wealth of information that will take years to analyze. Meanwhile, the many-ringed, many-mooned Saturn continues to present a fascinating spectacle.

Four views of Saturn, each taken when the planet has a different inclination relative to earth.

Lowell Observatory photograph



THE OUTERMOST PLANETS

Leaving behind us Saturn and its gorgeous retinue of rings and satellites, we proceed to far more remote areas of the solar system. Here we come upon two immense planets, Uranus and Neptune, and a tiny one, which is called Pluto.

Saturn was the outermost of the planets known to the ancient world. After astronomers had adopted the Copernican theory, which states that the planets revolve around the sun, they carefully calculated the orbit, or path, of Saturn. Until the eighteenth century it was taken for granted that no planet existed outside of this orbit. Perhaps astronomers would still be holding that belief if it had not been for a chance discovery made by a German-born English astronomer named William Herschel.

DISCOVERY OF URANUS

On the night of March 13, 1781, Herschel was observing the stars in the constellation Gemini with an 18-centimeter reflector telescope that he had just made. Suddenly he beheld what seemed to be a star shaped like a disk. Herschel was puzzled, since all true stars (with the exception of the sun) appear as mere points of light, even when viewed with the most powerful telescope. Observing the unknown body night after night, he noted that it changed its position among the stars. He then came to the conclusion that his "moving star" was a comet. He described it as such in a report that he sent to the Royal Society, the prestigious British scientific society that has been a center for the reading and discussion of scientific papers since the mid-17th century.

The supposed comet was carefully followed by astronomers. They noted that it followed an almost circular orbit far outside the orbit of Saturn. As time went on, they realized that the new body was a planet and hailed Herschel as its discoverer.

Herschel named the new planet Georgium Sidus (the Georgian Star) after the reigning monarch, George III. English as-

tronomers called the planet the Georgian Star until about 1850; to others it was known as Herschel. The name finally given to it—Uranus—was proposed by the German astronomer Johann Elert Bode, who pointed out that all the other planets had been named after ancient gods.

The mean distance of Uranus from the sun is 2,870,000,000 kilometers. It makes one complete revolution of the sun in 84.01 years. That is, its year is equal to about 84 earth years. The orbit is inclined to the plane of the ecliptic by less than one degree. The plane of its equator is almost at right angles to the plane of its orbit. Uranus has a diameter of about 50,100 kilometers at the equator and is about one fourth as dense as the earth. It rotates about its axis once every 24 hours.

Uranus can barely be made out by the naked eye on a moonless night. Through a telescope it appears as a slightly sea-green disk. There are a few noticeable markings including several faint belts, whose nature remains a mystery to astronomers. The atmosphere of Uranus is made up chiefly of methane. The maximum surface temperature of the planet is -180° Celsius.

Uranus has five moons, which have been given the poetic names of Ariel, Umbriel, Titania, Oberon, and Miranda. The last, which is the faintest and closest to the planet, was discovered in 1948 by the Dutch-born American astronomer Gerard P. Kuiper. The satellites revolve about Uranus in the plane of the planet's equator.

ENCIRCLED BY RINGS

In early 1977 astronomers discovered that Uranus is encircled by five rings. The finding, made by Dr. James Elliot of Cornell University's Center for Radiophysics and Space Research and his assistants, revealed "the first major structures in the solar system to be found since the discovery of the planet Pluto in 1930."

The rings lie 18,000 kilometers above the cloud tops of the planet. The entire

belt of five rings is about 7,000 kilometers wide. The rings range from about ten to 100 kilometers wide, and all, except perhaps the outermost, are circular. Like the rings of Saturn, these rings are thought to be made up of countless fragments that either failed to coalesce to form a satellite or else are the remains of what was once a satellite.

UNEXPECTED ORBIT

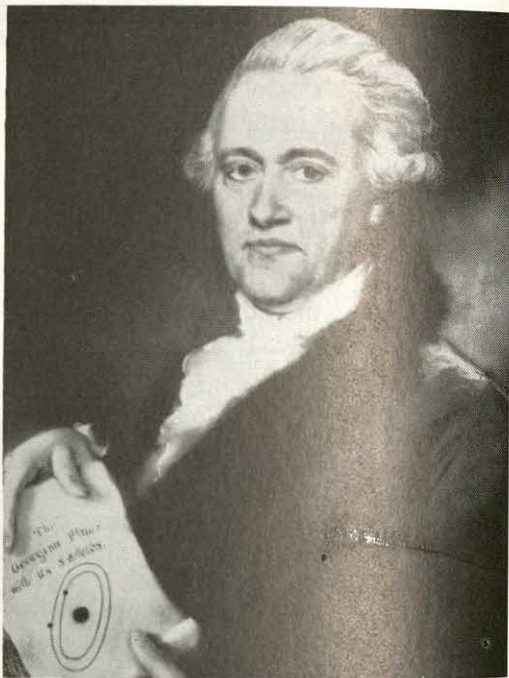
The neighbor of Uranus, huge Neptune, might still be unknown to us if Uranus had followed the orbit that the French astronomer Jean Baptiste Delambre mapped out for it in 1790. For a few years Delambre's tables were in reasonable agreement with the observed position of Uranus. As time went on, however, the discrepancy between the tables and the actual orbit of Uranus increased. Astronomers agreed on the need for new tables.

The French astronomer Alexis Bouvard undertook this task. His tables seemed quite accurate at first, but after a while they failed to agree with the actual positions of the planet. Bouvard became convinced that a planet lying beyond the orbit of Uranus must be responsible.

ANOTHER PLANET?

On July 3, 1841, John Couch Adams, then an undergraduate in St. John's College, Cambridge, entered the following memorandum in his notebook: "Formed a design, in the beginning of this week, of investigating as soon as possible after taking my degree, the irregularities in the motion of Uranus, which are as yet unaccounted for; in order to find whether they may be attributed to the action of an undiscovered planet beyond it, and if possible thence to determine approximately the elements of its orbit, etc., which would probably lead to its discovery."

By 1845 Adams had completed some ingenious calculations that showed that a heavenly body—probably a planet—beyond the orbit of Uranus was drawing the latter planet from its calculated course. The next step was to have some competent observer carefully search the



The Science Museum, London

Sir William Herschel, who, on March 13, 1781, discovered the planet Uranus.

region of the skies in which the planet would be found if Adams' calculations were correct. In September, 1845, therefore, he communicated his results to the eminent British Astronomer Royal, Sir George Biddell Airy, and asked him to undertake the search.

For a time Airy did nothing. According to one account, Adams failed to answer a crucial question that the Astronomer Royal had asked him. According to another, Airy did not go forward with the matter because he did not have a suitable star chart of the part of the sky indicated by Adams. It is possible, too, that he was not interested in the matter at first because Adams was unknown to him. Later, Airy bestirred himself. In July 1846 he asked the English astronomer James Challis to search for the planet. Challis observed Neptune on August 4, 1846, but he failed to recognize it.

DISCOVERY OF NEPTUNE

In the meantime the French astronomer Urbain-Jean-Joseph Leverrier had calculated the unknown planet's orbit, quite

independently of Adams. Leverrier sent a series of three memoranda to the French Academy on that subject—in November 1845 and in June and August 1846. Then he wrote to Johann G. Galle, chief assistant at the Berlin Observatory, enclosing his calculations and suggesting that Galle should look for the planet. The letter reached Galle on September 23, 1846. That same night he turned his telescope toward the quarter of the heavens suggested by Leverrier and found the planet less than one degree from the place where Leverrier said it would be.

The discovery of Neptune was made possible because the Berlin Observatory had just received a new and complete map of the stars in that region of the skies. The new planet was named Neptune, after the Roman god of the sea (identified with the Greek god Poseidon). Its discovery was a fine example of what has been called the "astronomy of the invisible": that is, the detecting of heavenly bodies, before they are actually observed in the skies, through the attraction they exert on known bodies.

There was a good deal of rather unpleasant controversy for a time concerning the matter of priority in the discovery of Neptune. It is now quite generally agreed that Adams and Leverrier deserve equal credit for the abstruse calculations that led to the discovery. To Galle goes the distinction of having been the first to identify the planet in the heavens.

The mean distance of Neptune from the sun is 4,500,000,000 kilometers. Its elliptical orbit is only slightly eccentric: that is, the sun is almost at the center of it. The plane of Neptune's equator is inclined by

about 29 degrees to the plane of the planet's orbit. It takes Neptune about 165 years to complete its rotation about the sun; its rotation period is 22 hours. The diameter of the planet is about 48,600 kilometers at the equator. Neptune is a little more than one half as dense as the earth.

GREENISH DISK OF NEPTUNE

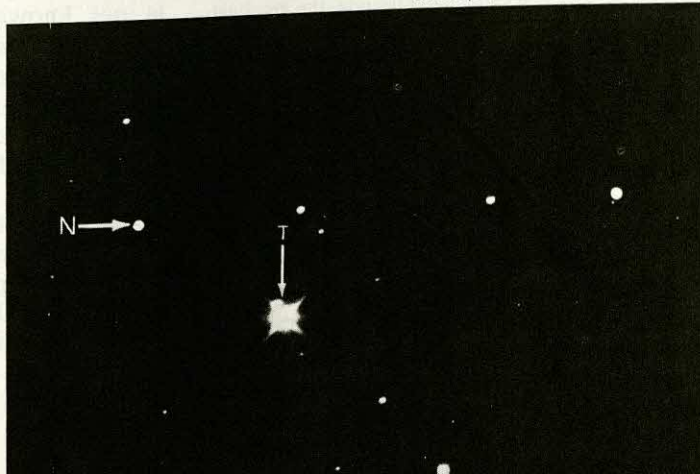
Like Uranus, Neptune shows a greenish disk. This is invisible to the naked eye but is plain enough when viewed through a telescope. The apparent magnitude of the planet at greatest brilliancy is 7.6. As a reflector of the sun's light, Neptune ranks high among the planets, surpassing even the bright planet Venus in this respect.

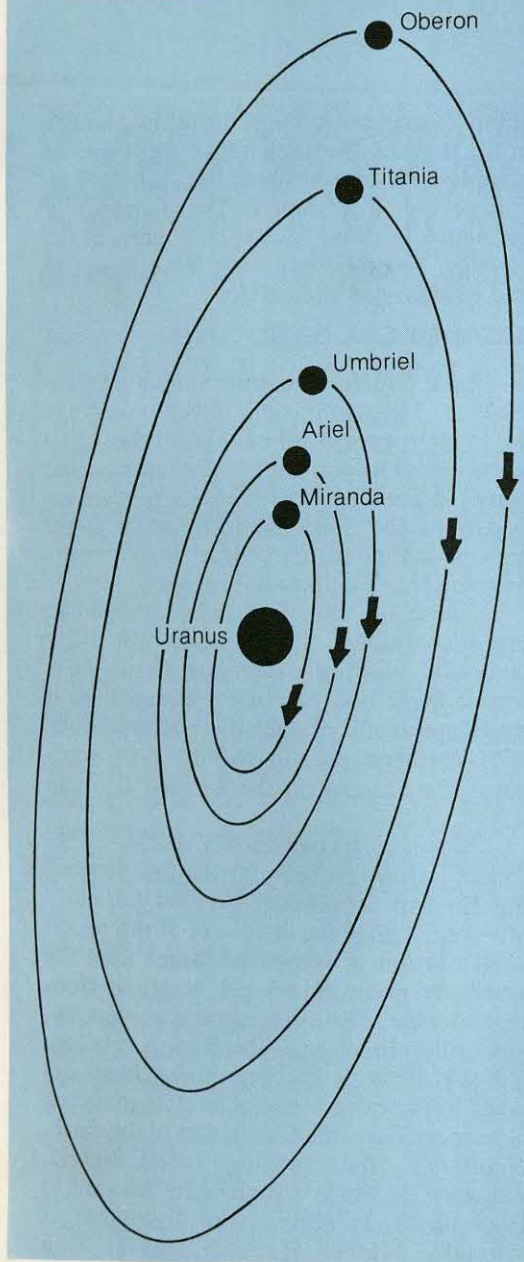
There are no clear markings on the planet's surface, though under the most favorable viewing conditions vague belts can be made out. Neptune's atmosphere is made up chiefly of methane and ammonia, with methane predominating. The maximum temperature at the surface is about -190° Celsius.

Neptune has two known satellites. One called Triton was seen for the first time by the English astronomer William Lassell a few weeks after the discovery of the planet itself. Triton is somewhat larger than the moon. It revolves around Neptune from east to west. This is retrograde motion, opposite the direction of Neptune's rotation. Triton's orbit is inclined about forty degrees to the orbit of the planet. Its mass is a little more than one fiftieth that of the earth. Neptune's other satellite, called Nereid, was seen for the first time in the heavens in 1949 by G. P. Kuiper, the discoverer of Miranda. Nereid is much smaller than

G. P. Kuiper, Lunar and Planetary Laboratory

The planet Neptune shown with its two satellites: Triton, which is very close to the planet, and the smaller, more distant Nereid.





Uranus with its five satellites. Miranda, the smallest and closest to the planet, was discovered in 1948 by the Dutch-American astronomer G. P. Kuiper.

Triton and is far more distant from Neptune. It is inclined about five degrees to the ecliptic; its motion is direct (from west to east).

DISCOVERY OF PLUTO

Even after Neptune had been discovered and its orbit calculated, the planet

Uranus did not follow the course that had been mapped out for it. Could the planet's eccentric orbit be due to still another planet lying beyond the orbit of Neptune? A number of astronomers, including Percival Lowell, director of the Lowell Observatory, near Flagstaff, Arizona, decided to investigate the matter. By 1905, Lowell had made preliminary computations of the probable position of what he called the trans-Neptunian (beyond Neptune) planet. Under his direction the staff members at the observatory began to search for it.

In 1915, Lowell published his *Memoir on a Trans-Neptunian Planet*, in which he presented his mathematical calculations of the planet's position. He believed that it would be found in either one of two areas in the sky. The search for the trans-Neptunian planet continued; years went by. At last on February 18, 1930, Clyde W. Tombaugh, an assistant on the Lowell Observatory staff, found what seemed to be the long-sought planet as he studied a series of photographs he had taken of the skies. Further observations were made, and finally, on March 13, the Lowell Observatory officially announced the discovery of the new planet. It was named Pluto, after the Greek god of the underworld, because of its position in the distant part of the solar system.

The planet Pluto is considerably smaller than the earth. Its diameter is about 6,000 kilometers at the equator. In 1978, the discovery of a satellite revolving around Pluto was announced. This second body is thought to be only 2 to 3 times smaller than Pluto itself. Pluto appears as a faint, yellowish point of light of the fifteenth apparent magnitude. Its period of rotation is not known. Its period of revolution around the sun is 248.43 years. The orbit of Pluto is more inclined to the ecliptic than that of any other planet. The orbit is quite eccentric; that is, the sun is relatively far removed from the center of the orbit. Hence, when Pluto is at perihelion, or its position nearest the sun, it is nearer to the earth than is Neptune. The mean distance of the planet from the sun is about 6,000,000,000 kilometers. The surface temperature of Pluto is about -220° Celsius.

Recent evidence indicates that Pluto is covered with frozen methane. Estimates of the planet's size have been based on its brightness as a point of light. If, however, the planet is covered with methane ice, it could be much smaller than previously believed and still shine as brightly. Some astronomers suspect that Pluto might be a former satellite of Neptune.

MORE PLANETS?

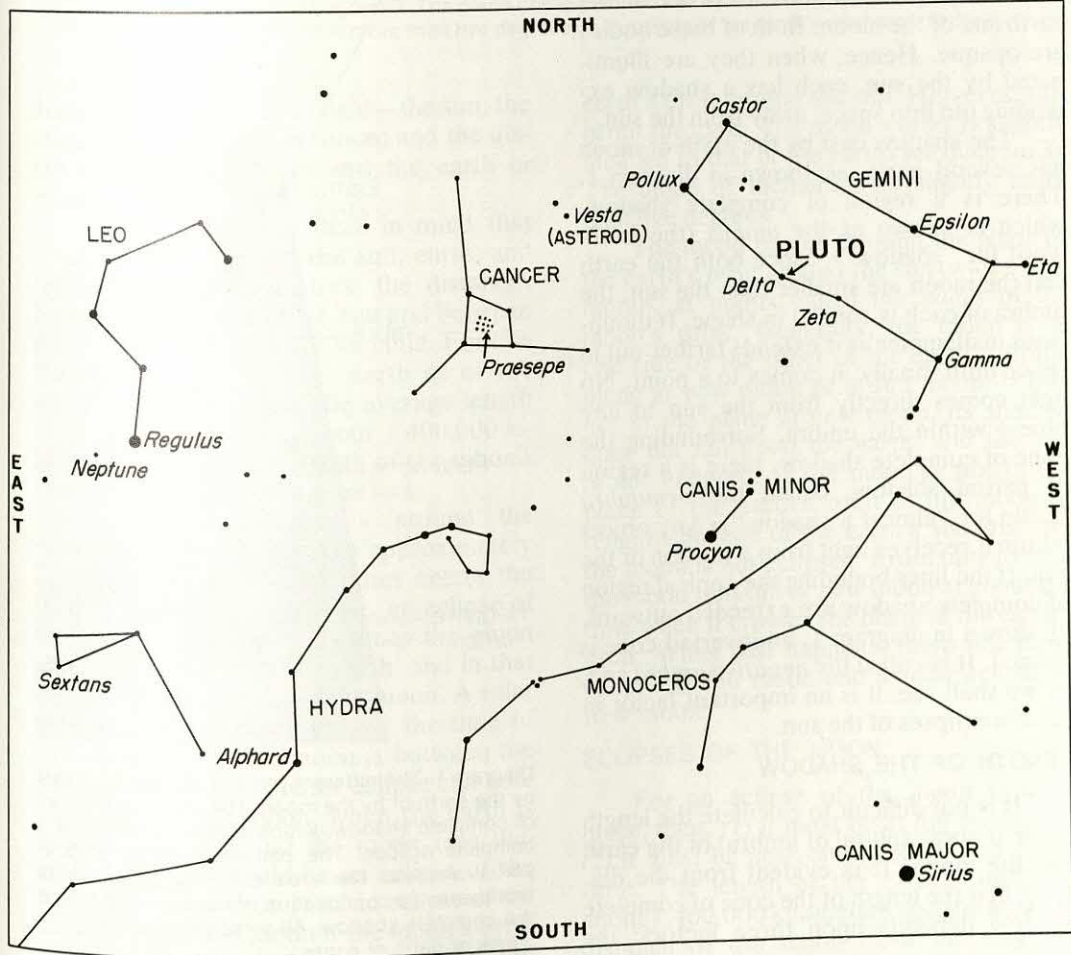
Have we now discovered all the planets that circle the sun? It would be a rash astronomer indeed who would answer the

A drawing showing the position of Pluto about the time of its discovery. Careful study of a series of photos had revealed the long-sought planet X as a moving dot slightly west of Delta.

question with an emphatic "yes"—particularly after the startling discoveries of the rings around Uranus and Jupiter. Some authorities still maintain that discrepancies in the calculated orbits of the outermost planets may be due to yet another—still undiscovered—planet.

Much more information about the outer solar system may be gathered in the late 1980's. Following the Saturn flyby, Voyager 2 continued outward toward a planned close encounter with Uranus in 1986. The probe is then scheduled to sweep by distant Neptune in 1989. There it may confirm the suspicion that a ring system encircles the planet. This grand journey, then, may provide us with our first good look at the outer solar system.

H. S. Rice, American Museum of Natural History



ECLIPSES

An eclipse of the sun or moon is truly an awe-inspiring spectacle. The word "eclipse" comes from the Greek word *ekleipsis*, meaning "forsaking," or "abandonment," indicating how the ancients dreaded this celestial drama. As the sun or moon disappeared from view, it seemed indeed to be deserting mankind. Eclipses, like comets, were held to be portents of war, pestilence, the death of princes, or even the end of the world. To this day, certain primitive peoples come to the aid of the sun or moon, as it is being eclipsed, with solemn rites and loud entreaties.

We know now that there is a perfectly logical explanation of eclipses: they are caused by the enormous shadows of the earth and of the moon. Both of these bodies are opaque. Hence, when they are illuminated by the sun, each has a shadow extending out into space, away from the sun.

The shadow cast by the earth or moon has several parts, as shown in diagram 1. There is a region of complete shadow, which is known as the *umbra* (the Latin word for "shadow"). Since both the earth and the moon are smaller than the sun, the umbra of each is conical in shape. It diminishes in diameter as it extends farther out in space until, finally, it comes to a point. No light comes directly from the sun to any object within the umbra. Surrounding the cone of complete shadow, there is a region of partial shadow, called the *penumbra* (Latin for "almost a shadow"). Any object within it receives light from a portion of the sun. If the lines bounding the conical region of complete shadow are extended outward, as shown in diagram 1, an inverted cone is formed. It is called the *negative umbra* and, as we shall see, it is an important factor in certain eclipses of the sun.

LENGTH OF THE SHADOW

It is not difficult to calculate the length of the umbrae (plural of umbra) of the earth and the moon. It is evident from the diagram that the length of the cone of complete shadow depends upon three factors: the

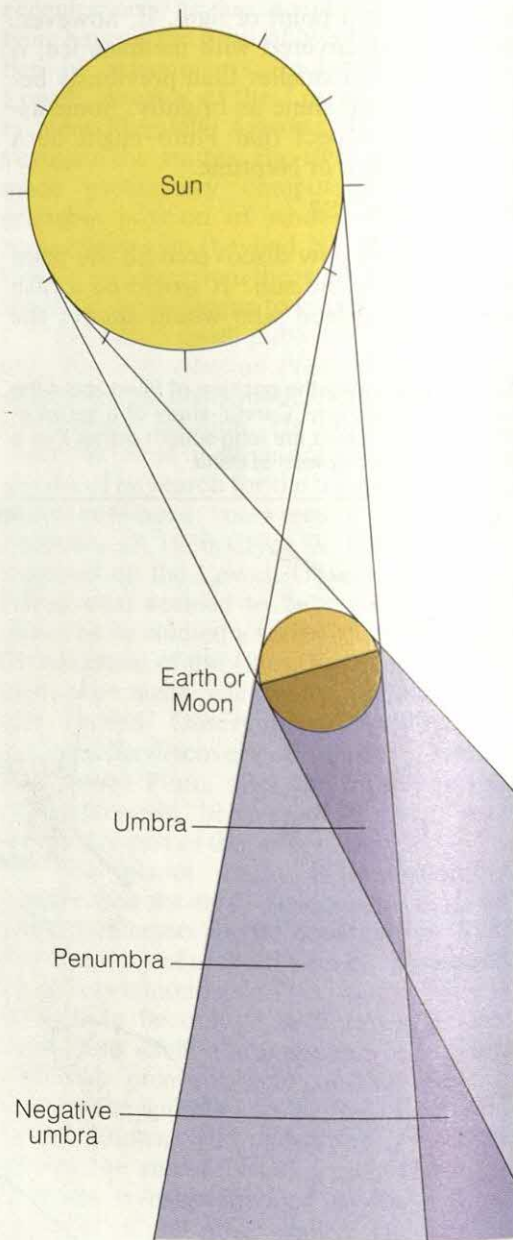


Diagram 1. The different parts of the shadows cast by the earth or by the moon. The umbra is the area of complete shadow, which appears on earth as a complete eclipse. The penumbra is the area of partial shadow. The area of the negative umbra represents the continuation of the lines that bound the complete shadow. All three shadow areas are cones or parts of cones.

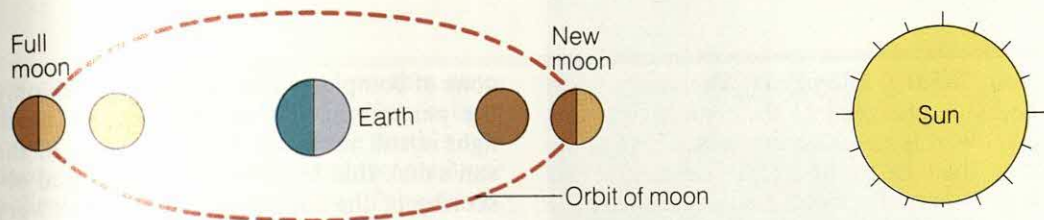


Diagram 2. Two of the phases of the moon: new moon and full. The inner circles along the moon's orbit show the phases as viewed from earth.

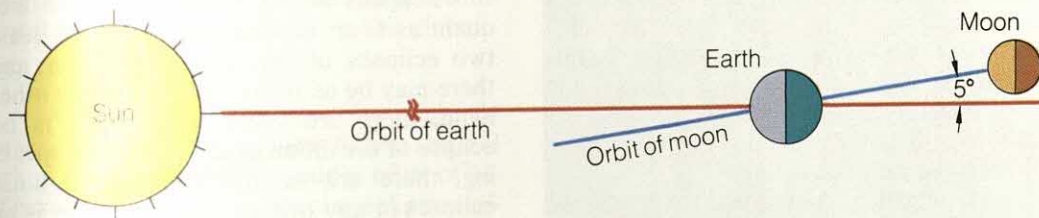


Diagram 3. The plane of the moon's orbit around the earth is slightly inclined (a little more than five degrees) to the plane of the earth's orbit around the sun.

diameter of the source of light—the sun; the diameter of the earth or moon; and the distance between the sun and the earth or moon.

It is important to bear in mind that while the diameters of the sun, earth, and moon are constant factors, the distances between the earth and the sun and between the moon and the sun are variable. For this reason, the umbra of the earth or of the moon varies in length. The average length of the earth's umbra is about 1,400,000 kilometers; the average length of the moon's umbra, about 375,000 kilometers.

As the moon travels around the earth—a journey that takes approximately a month—the earth sometimes enters the moon's shadow. In that case, an eclipse of the sun takes place. Sometimes the moon enters the shadow of the earth, and in that case there is an eclipse of the moon. A solar eclipse can take place only at the time of new moon, when the moon is between the sun and the earth. A lunar eclipse can take place only at full moon, when the earth is between the sun and the moon (see diagram 2). There would be an eclipse of the sun at every new moon and an eclipse of the moon at every full moon if the moon's orbit were in exactly the same plane as the

earth's orbit around the sun. However, this is not the case. The moon's orbit is slightly inclined to that of the earth (see diagram 3). The angle of inclination is slightly more than five degrees.

The moon passes through the plane of the earth's orbit around the sun twice every month, at points called the *nodes* of the moon's orbit. Generally, the moon is at one side or the other of the earth's orbital plane at new moon or at full moon. If it is not in the plane at new moon, its shadow does not fall upon the earth, and the sun is not eclipsed. If the moon is not in the plane of the earth's orbit at full moon, it remains outside of the earth's shadow, and the moon is not eclipsed. From time to time the moon is at full or new moon at about the time when it crosses the plane of the earth's orbit. When that happens, there is a lunar eclipse at full moon and a solar eclipse at new moon.

ECLIPSES OF THE MOON

For an eclipse of the moon to take place, then, (1) it must be at full moon and (2) it must be near one of the nodes of its orbit. The length of the earth's umbra is about 1,400,000 kilometers and the average distance of the moon from the earth is

about 385,000 kilometers. Therefore, when the moon plunges into the cone of complete shadow, it is much nearer to the base of the cone than to its tip. The diameter of the cone, where the moon passes through it, is about two and one-half times the diameter of the moon.

If the path of the moon happens to pass through the center of the shadow, the moon may remain totally eclipsed for about an hour. The shadow may cover part of it for about two hours. A lunar eclipse begins when the moon enters the penumbra and ends when it leaves the penumbra. There is little significant darkening, however, until the moon enters the umbra.

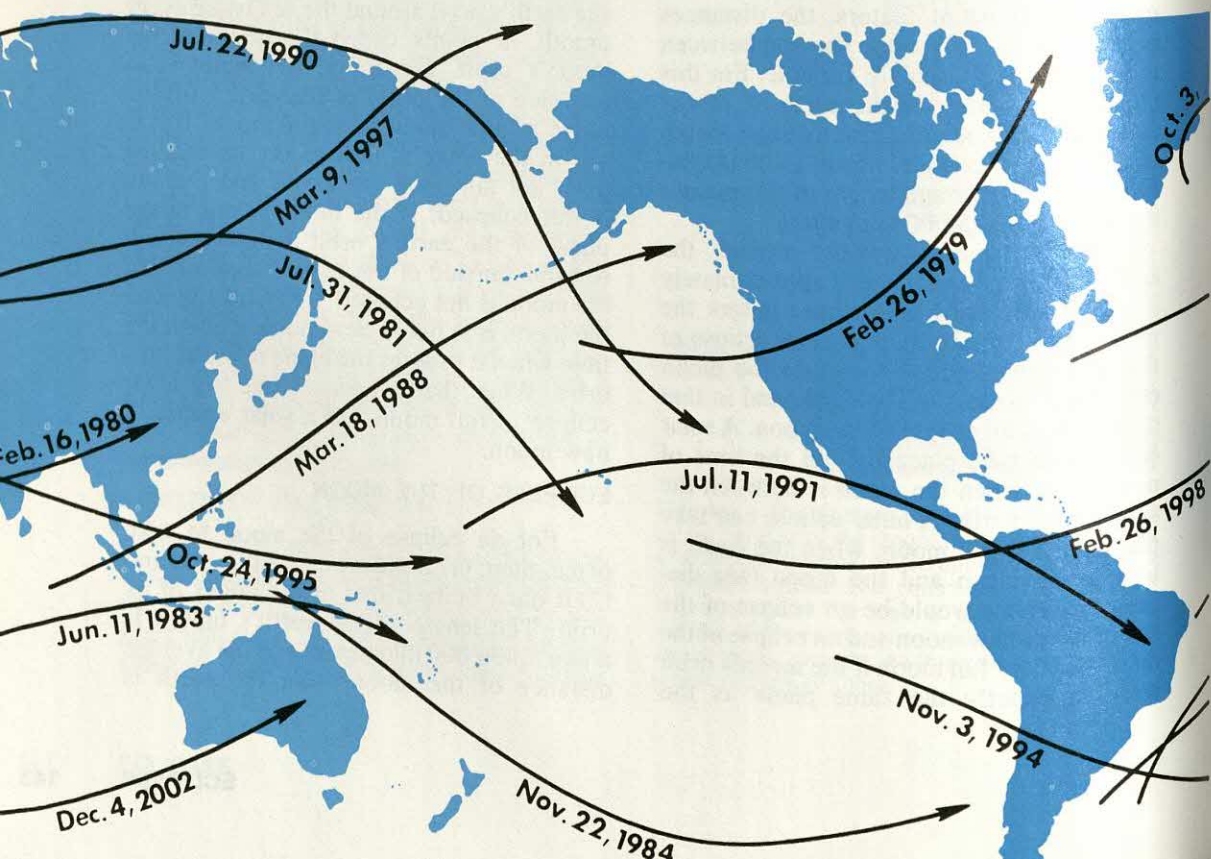
If the path of the moon takes it near the edge of the shadow, the total phase of its eclipse may last only a few minutes. If the moon's path is such that only a portion of its disk, and not the whole of it, enters into the conical shadow of the umbra, the eclipse is partial and not total. Sometimes the moon in its path passes, not through the

cone of complete shadow, but through only the penumbra. If this happens, so much light is still received from a portion of the sun's disk that there will be no marked obscuring of the moon unless it passes very close to the true shadow. The moon is usually not altogether lost to view even in the midst of a total eclipse. It shines with a strange copper-colored glow.

Eclipses of the moon are not so frequent as solar eclipses. There are at least two eclipses of the sun every year, and there may be as many as five. On the other hand, there are years when there is no eclipse of the moon at all. Generally speaking, there are no more than two lunar eclipses in any one year. But if there is an eclipse of the moon on one of the first days of the year, there may be a third eclipse in December.

The statement that eclipses of the sun are more frequent than eclipses of the moon may seem to run counter to our experience. After all, we know that people living in a given region will see more eclipses of the moon than eclipses of the sun within a specific period of time. But, as a matter of fact,

Paths of important total eclipses of the sun in the 1970's and 1980's. Arrows show the direction of movement.



no contradiction is involved in this case. The moon's shadow in a solar eclipse covers a small part only of the earth's surface, whereas the earth's shadow in a lunar eclipse covers the entire face of the moon. Every eclipse of the moon, therefore, is visible over that half of the earth that is in darkness (that is, where it is night). However, the regions from which any particular eclipse of the sun can be seen lie in a comparatively narrow track across the globe.

ECLIPSES OF THE SUN

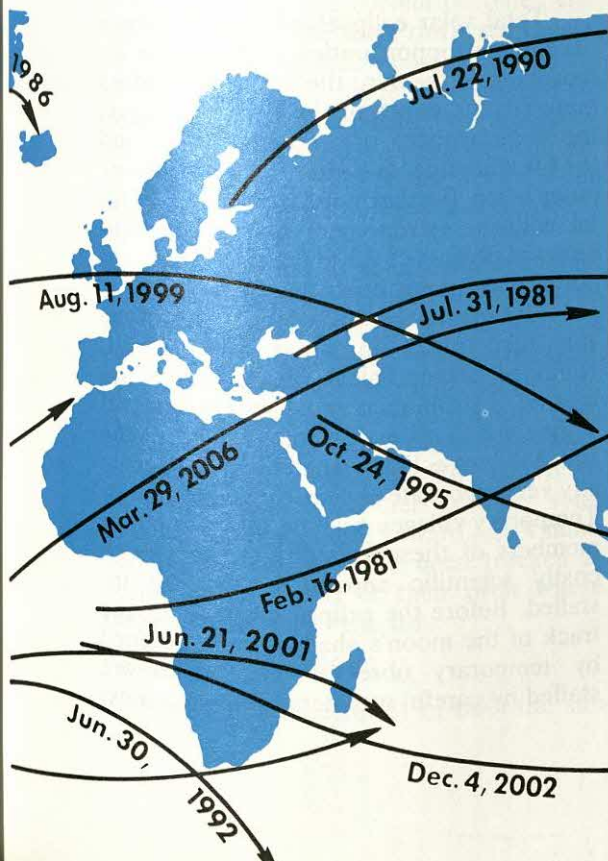
A solar eclipse can take place only (1) when the moon is at new moon and (2) when it is near one of the nodes of its orbit. There are three kinds of solar eclipses: *total*, *annular*, and *partial*.

We have seen that the average length of the moon's shadow is about 375,000 kilometers. It never extends beyond about 380,000 kilometers. The average distance of the earth from the moon, however, is about 385,000 kilometers, so that in general the moon's umbra—its true shadow—is not long enough to reach the earth. At times, however, the moon is only about 356,500



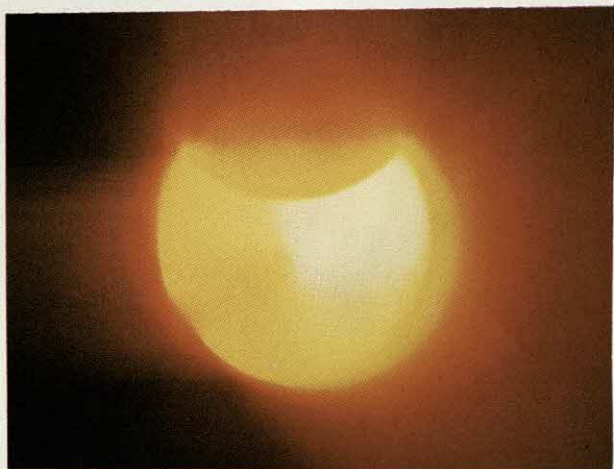
Gamma

A total solar eclipse. The aberrations above the house were caused by the reflection of the camera.



kilometers from the earth's surface. The true shadow then falls upon a small part of the earth's surface, causing a total eclipse of the sun over an area that cannot exceed 270 kilometers in diameter. At other times, the moon may be more than 406,000 kilometers from the earth. If it is then interposed between the sun and the earth, its negative umbra will partially obscure a small area of the earth's surface, causing an annular eclipse of the sun at that point. In an annular eclipse, the sun's rim appears as a "ring" of light around the dark moon. (*Annulus* means "ring" in Latin.)

Around the area where there is a total eclipse or an annular eclipse of the sun, there is always a much larger area where there is a partial eclipse. This area is in the moon's penumbra. It generally extends for more than 3,000 kilometers of the earth's surface on each side of the path where the total eclipse can be seen. Sometimes the area of partial eclipse extends nearly 5,000 kilometers on each side of the path of totality.



This page and next: four successive phases of the 1973 total eclipse, as seen from an airplane.

The moon's shadow passes along this path at great speed—about 1,700 kilometers an hour. The longest period of total eclipse, under the most favorable conditions, is about seven and one-half minutes.

Few spectacles in the heavens are as startling as a solar eclipse. The approach of a total solar eclipse is particularly impressive and, to some persons, alarming. The sky darkens; birds fly to shelter; other animals may show signs of alarm. The darkness rapidly increases. Finally the dark shadow of the moon, like a vast thundercloud, advances with awe-inspiring rapidity from the western horizon and covers the land. Usually, just before the last rays of the sun are obscured, swiftly moving bands of light and shade are observed. (They are probably due to uneven refraction in the atmosphere.) Then the day becomes like night. As the eye becomes accustomed to the darkness, surrounding objects seem to have an eerie appearance. No wonder eclipses were once regarded with dread.

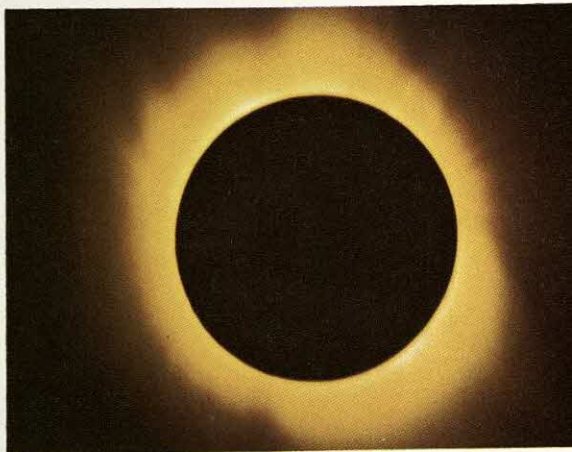
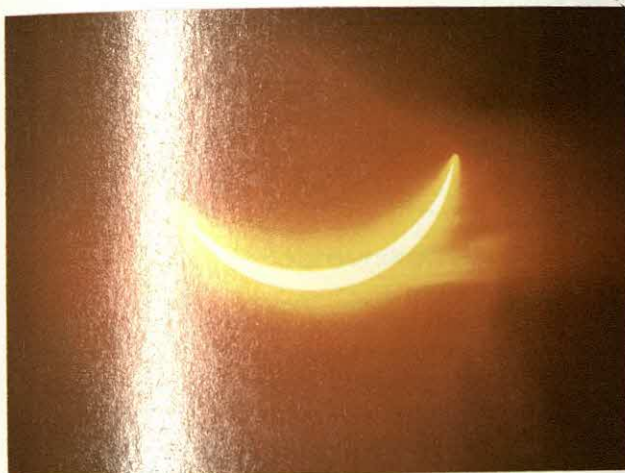
As the eclipse approaches totality, the sun is seen as a very narrow crescent of brilliant light. The crescent then becomes a curved line and finally breaks off into irregular beads of light known as Bailey's beads. The beads are caused by irregularities in the outline of the moon as seen in silhouette. The irregularities are due to lunar mountains and valleys. The sun's rays pass through the valleys, but are obscured by the

mountains. Even at totality, the atmosphere that surrounds the sun extends so far out in space that it is never completely covered by the moon. We see the eclipsed sun surrounded by a beautiful glow. While observing a solar eclipse, remember that the eyes must at all times be protected from the direct rays of the sun.

OPPORTUNITY FOR STUDY

Total solar eclipses offer astronomers exceptional opportunities to study the atmosphere of the sun; the distribution of its material; the depth of the reversing layer; the chromosphere, or sun's inner layer; and the breathtaking splendor of the corona, or outer layer. Furthermore, at the time of total eclipse, astronomers can photograph celestial objects close to the sun. This cannot be done at any other time.

Unfortunately, few total eclipses of the sun take place in areas where well equipped astronomical observatories are located. Astronomers must generally travel if they are to observe a total eclipse of the sun. Modern eclipse expeditions are generally very elaborate and costly undertakings. Temporary villages must be set up to house members of these expeditions, and much costly scientific apparatus must be installed. Before the eclipse takes place, the track of the moon's shadow must be lined by temporary observatories, which are staffed by careful watchers. The work must



all photos, Gamma

be carefully planned and timed, since an eclipse lasts for only a few minutes.

The simplest observations made during a total eclipse of the sun are the precise moments of each of the so-called contacts. These give detailed information about the moon's position and motion. The first contact takes place when the moon first encroaches on the sun's disk. In the second contact, the sun disappears behind the moon, and the solar corona becomes visible. In the third, the sun's rim is seen again. In the fourth, the moon passes completely off the sun's disk. If the times of contact are observed at widely separated stations along the path of the eclipse, scientists can determine the exact distances from certain points on the earth's surface to other points far removed.

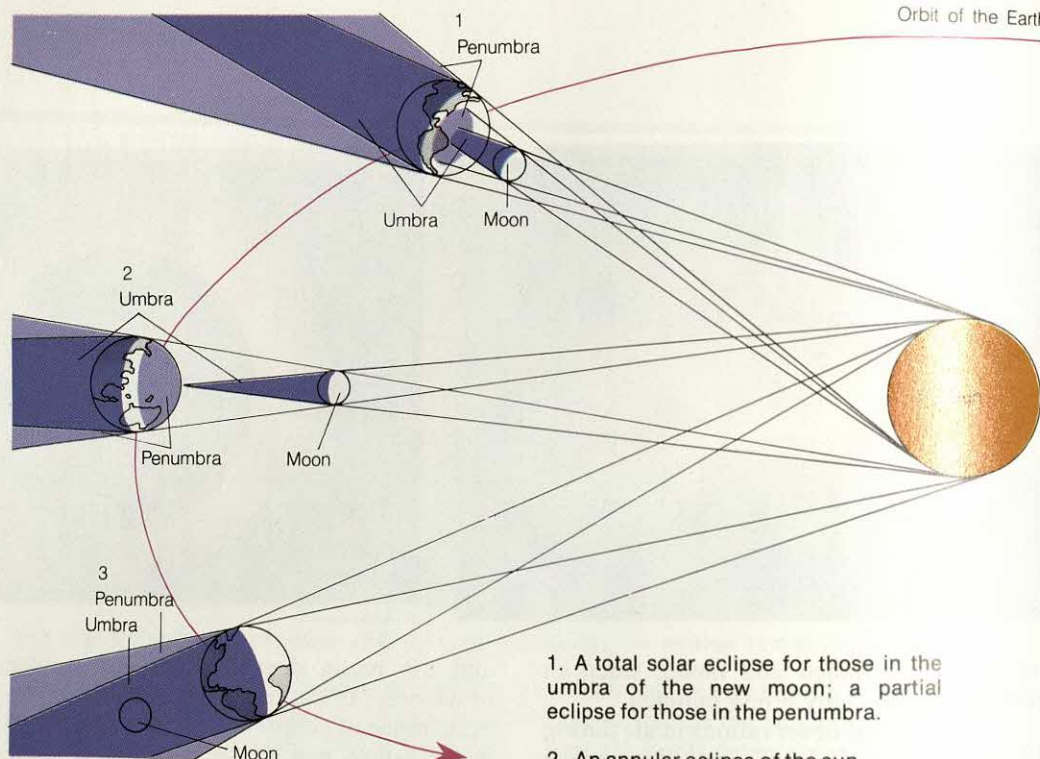
In the period of total eclipse, certain observers search for possible undiscovered planets within the orbit of Mercury, the planet closest to the sun. At any other time, the brilliance of the sun's light makes it impossible to spot a planet near the sun. Thus far, no new planet within the orbit of Mercury has been discovered. The search for hitherto unknown comets during solar eclipses has been more rewarding. A number of new comets have been tracked at perihelion—the part of their orbit where they pass closest to the sun.

PREDICTING ECLIPSES

Inasmuch as eclipses of both the sun

and the moon depend upon the regular movements of the sun, earth, and moon, the occurrence of eclipses, both in the past and in the future, can be calculated with great accuracy. In the 1880's, for example, the Austrian astronomer Theodor Oppolzer published a book called the *Canon of Eclipses*, in which he gave a table of 8,000 solar eclipses and 5,200 lunar eclipses taking place between 1207 B.C. and A.D. 2162. His *Canon* has now been extended to 2510 A.D. and revised with the aid of electronic computers. In the case of solar eclipses, he indicated the areas of the earth from which the eclipses would be visible. By consulting Oppolzer's tables, we find, for instance, that on February 16, 1980, a total eclipse of the sun lasting four minutes and eight seconds will take place in various parts of Africa.

Predictions of eclipses go back to antiquity. Perhaps the most famous of these early predictions was that of Thales, a philosopher of Miletus, who died in the year 546 B.C. The Greek historian Herodotus has given the following dramatic account of the prediction: "There was war between the Lydians and the Medes . . . In an encounter which happened in the sixth year [of the war] it chanced that the day was turned into night. Thales of Miletus had foretold this loss of daylight to the Ionians [Miletus was in the ancient district of Ionia], fixing it within the year in which the change did indeed happen. So when the



1. A total solar eclipse for those in the umbra of the new moon; a partial eclipse for those in the penumbra.

2. An annular eclipse of the sun.

3. A total eclipse of the moon.

Lydians and Medes saw the day turned to night they ceased from fighting, and both were the more zealous to make peace." This eclipse has been identified with the one that took place on May 28, 585 B.C.

Thales' prediction was based on a remarkable discovery made by Chaldean astronomers long before his time. They had noted that eclipses of the sun and moon occur in series and that a definite period of time elapses between one eclipse of a series and the following eclipse. By calculating the time that had passed since the last solar eclipse in a series of this sort, Thales could predict the next eclipse.

SERIES OF ECLIPSES

The interval between an eclipse of the sun or moon and the next one in a given series is called a *saros*. Each saros is 18 years and $11\frac{1}{3}$ days long (or 18 years and $10\frac{1}{3}$ days long if there are five leap years instead of four in a saros). While a given series of solar eclipses may run through 70 saroses and last for 1,250 years, there are only 48 or 49 saroses in the average series of lunar eclipses. A given series of lunar eclipses lasts somewhat under 900 years.

Several eclipse series, solar and lunar, run their courses at the same time. If an eclipse in Series A took place, say, in the first week of May, 1970, the next eclipse of that same series will occur in May, 1988. In another series, which we shall call Series B, an eclipse may occur, say, in the last week of June, 1971. The next will be in July 1989.

The first eclipse in a series of solar eclipses is partial, the moon encroaching but slightly on the sun's disk. At the next eclipse, the moon obscures a somewhat larger area of the sun. Next time, the eclipse, though still partial, is more pronounced. It becomes more extensive with each passing saros. Ultimately, the partial eclipses are followed by a series of annular and total eclipses. In these, as the moon passes across the sun's disk, the moon either covers all but the outer rim (an annular eclipse), or it obscures the sun altogether (a total eclipse). Then there follow a succession of partial eclipses, each one obscuring a smaller area of the sun's disk than the one before. In the last eclipse of this particular series, the bright disk of the sun is only slightly obscured.

As Comet West moved away from the sun, it became very bright with an unusually long two-part tail. This photo shows it against a dawn sky on March 6, 1976.

COMETS

by Nicholas T. Bobrovnikoff

As the ancients gazed at the night skies, they were occasionally startled to see strange celestial objects intruding upon the familiar pattern of stars, moon, and planets. These mysterious apparitions looked like fuzzy stars with long trains of light, moving from one constellation to another and cutting across the paths of the planets at every conceivable angle. The trains of light suggested a woman's tresses; hence the celestial intruders came to be known as "long-haired stars." We now realize that these so-called stars were comets, a name derived from the Greek word *kometes* for "long-haired."

A bright comet was a terrifying spectacle in antiquity and for a long time thereafter. It was thought to foreshadow some dire catastrophe—plague, or famine, or war, or perhaps the death of a ruler. In Shakespeare's *Julius Caesar*, Calpurnia, Caesar's wife, warns him not to venture forth on the fatal ides of March because she has beheld a comet. Says she:

"When beggars die, there are no comets seen; The Heavens themselves blaze forth the death of princes."

Of course no intelligent person believes today that comets are messengers of doom. We realize that they are bona fide members of the solar system and that their coming is no more portentous than the appearance of the first stars as twilight deepens.

HOW COMETS APPEAR

When comets are first discovered in the heavens, they usually appear as faint, diffuse bodies with a slight condensation, or denser section, toward the middle. This



Photograph by George Briggs; courtesy of Astronomy magazine

denser part, which sometimes looks like a tiny star, is known as the *nucleus*. The nebulous, or veil-like, region around it is the *coma*. Nucleus and coma together form the *head* of a comet. Most comets show no particular change in appearance.

In a certain number of cases, however, a spectacular transformation takes place as the comet approaches the sun. The coma changes from a diffuse round mass to sharply defined layers, called envelopes. Nebulous matter streams away from the comet's head in the direction opposite to the sun and forms an immense tail. Most comets of this type have only one tail. A very few have two or more. The bright comet of 1744 had six in all. Some comets occasionally also have forward spikes. As a comet recedes from the sun, the tail (or tails) can no longer be seen, the coma becomes diffuse again, and in the great majority of cases the comet itself disappears from view.

ORIGIN AND STRUCTURE

How do comets originate? According to one theory, they represent celestial building blocks left over after the formation of the planets. According to another, they are remnants of shattered worlds. All this is pure conjecture; so are the various theories that attempt to explain how comets are launched on their journey around the sun. One theory, proposed by the Dutch astronomer J. H. Oort in 1950, holds that there is a vast storehouse of comets—as many as 100,000,000,000, perhaps—in the icy reaches beyond the farthest planetary orbit. According to this theory, a given comet would normally remain entirely inactive in the “deep freeze” of space unless the passage of a star disturbed it. It then would swing into the sphere of gravitational attraction of one of the major planets, such as Jupiter or Saturn, and would revolve around the sun a few hundred or few thousand times until its substance would be depleted and it would disintegrate.

We are on more solid ground when we try to analyze the structure and composition of comets. The general belief is that the nucleus consists of a vast number of small solid bodies held together by mutual attraction. The nuclei of certain comets that have come close to the earth have been measured with considerable precision. The tail of the great comet of 1861 stretched across two-thirds of the sky, and the comet was bright enough to cause shadows to be cast on the earth. Yet it had a nucleus less than 160 kilometers in diameter. When Pons-Winnecke's comet came within 6,500,000 kilometers of the earth in 1927, its nucleus was found to have a diameter no greater than one or two kilometers.

As the nucleus of a comet approaches the sun, the solar heat vaporizes the material on the outer surface of the nucleus. Escaping gases, carrying fine dust with them, diffuse into the coma. They are then swept away by the force of the sun's radiation to form the tail. The gases and the dust they transport are illuminated partly by reflected sunlight. In part, too, they become luminous because they absorb ultraviolet light

and then give it out in the form of visible light.

The tail, which, as we have seen, is on the opposite side of the comet from the sun, increases in breadth as the distance from the head increases. The tail does not form an exact continuation of the imaginary line between the sun and the comet's head. The greater the distance from the comet's head, the more the gases and dust that make up the tail lag behind. Hence, the tail often has the shape of a curved horn, with the tip of the horn at the head.

When the comet turns away from the sun, the material that formed the tail is swept off into space. Obviously, in time, comets must gradually lose all their substance, unless it can be replenished by dust and by gas molecules swept up in the course of the journey through space.

Right: elliptical orbit (blue line) of Halley's comet, which will next be visible in 1986.

Below: possible orbits of comets. Periodic comets move in an ellipse (dotted line). Others move along open curves, either parabolic (dot-and-dash line) or hyperbolic (dashed line). Still others (solid line) disintegrate in the solar system.



COMPOSITION

The spectrograph has revealed that comets contain various gases, including cyanogen (CN), carbon (C_2), carbon monoxide (CO), nitrogen (N_2), hydroxyl (OH), and nitrogen hydride (NH). The known presence of highly poisonous gases, such as cyanogen and carbon monoxide, in comets gave rise to a certain uneasiness in 1910, when Halley's comet passed between the earth and the sun. Astronomers had announced that our planet was certain to pass through at least a part of the comet's tail. Would not the earth's inhabitants be subjected to gas poisoning? After the comet had passed, it was realized that such fears had been groundless. The gases in the tail did not produce the slightest effect as they swept over the surface of the earth.

Undoubtedly the earth has passed through the tails of comets many times and, as far as we know, without harm to living

organisms. The reason is that there are too few molecules in the tails to contaminate the earth's atmosphere. It has been calculated that the best vacuum obtainable in the laboratory contains millions of times more matter per unit of volume than does the tail of a comet.

SMALL AMOUNT OF MATTER

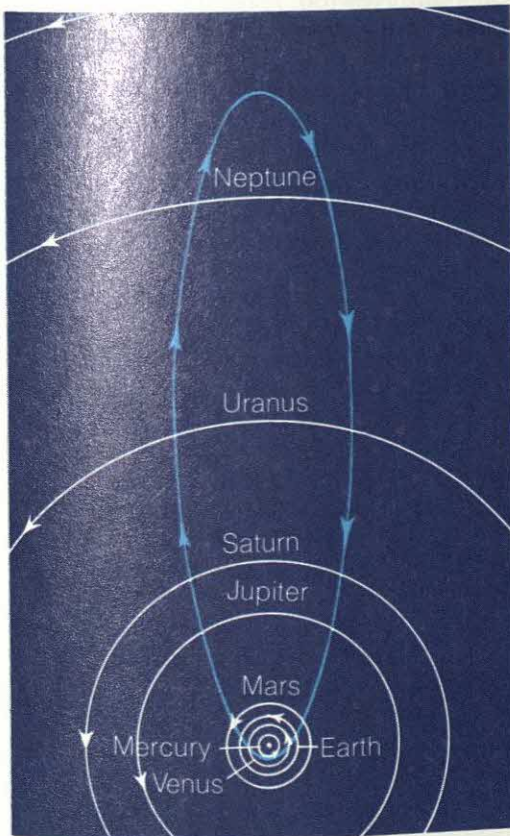
The total amount of matter in a comet, nucleus and all, is so small that it cannot be measured by the attraction that the comet exerts on the planets or their satellites. It has been discovered by indirect methods that the entire mass of Halley's comet cannot be more than 1/1,000,000,000th that of the earth.

But, although the mass of a comet is so small, its size may be very great. The head of the great comet of 1811 was considerably larger than the sun. As a rule the head of even a very small comet has a diameter larger than the earth's. The tail may stretch over millions of kilometers, sometimes reaching several hundred million kilometers.

Every time a comet approaches the sun, the strong gravitational pull exerted by that huge body subjects the comet to a tremendous strain. As a result, the comet may be broken up into two or more smaller bodies. That is what happened to Biela's comet, which split into two comets in the winter of 1845-46. The two were next observed in 1852 traveling not far apart in the old orbit. They were no longer visible at the predicted returns of 1859 and 1866. On what would have been the next approach of the comets to the sun—in 1872—a dazzling meteor shower was seen. The comets had disintegrated. We know that other meteor showers, including the Perseids in August and the Leonids in November, have been due to the disintegration of comets.

DIRTY SNOWBALL

The American astronomer Fred L. Whipple, of the Harvard Observatory, proposed a novel theory of the nature of comets—the so-called icy conglomerate, or dirty snowball, theory—in the early part of the 1950's. Dr. Whipple maintains that



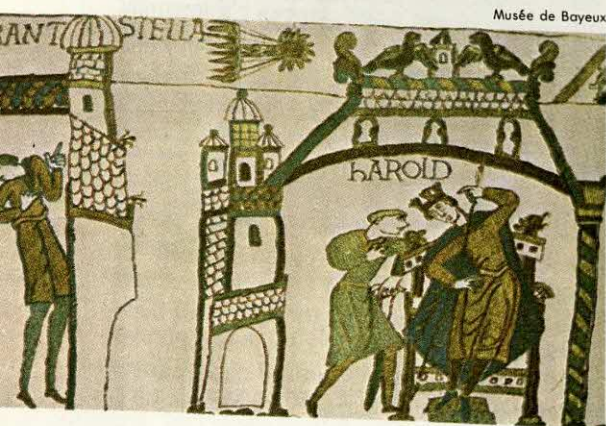


NASA

January 1974 photo of Comet Kohoutek. This photo was artificially colored to show brightness levels and the comet's hydrogen halo.

from 70 to 80 per cent of the mass of a comet is made up of icy particles, consisting of compounds of hydrogen with heavier elements, particularly carbon, nitrogen, and oxygen. These compounds might be methane (CH_4), ammonia (NH_3), and water (H_2O). As a comet approaches the sun, according to Whipple, the ice particles on the outer surface of the comet sublimates, or passes directly from the solid to the gaseous state. The resulting gases, together with fine dust particles, would form the tail. The remaining 20 to 30 per cent of the comet's mass—consisting of compounds of the heavier elements—do not vaporize appreciably during the lifetime of the comet.

Halley's comet is depicted in this detail of the 11th century Bayeux tapestry, which commemorates the Norman conquest of England in 1066.



Musée de Bayeux

These comparatively heavy compounds are the particles that produce spectacular meteoritic showers when the comet finally disintegrates.

MOTION OF COMETS

The motion of the comets baffled astronomers for a long time. Although the apparent motion of the planets is very complicated, it shows much regularity, which was evident even to the ancients. Comets, on the other hand, emerge into view with a flourish and then disappear for years on end. They traverse all regions of the sky at various angles to the plane of the solar system. No wonder that until the time of Tycho Brahe, a late 16th-century astronomer, comets were thought to be phenomena of the earth's upper atmosphere, like the aurora borealis, or northern lights.

It was not until 1705 that the true nature of cometary motion was established. In that year Edmund Halley, a friend of Sir Isaac Newton, applied the law of gravitation to the observations of a number of comets. He found that they traveled in space in accordance with that law. He noted, too, that the comets of 1682, 1607, 1531, and 1456 had moved in much the same way. He came to the conclusion, therefore, that these supposedly different comets were really one and the same body, which reappeared every seventy-five or seventy-six years. Halley predicted that the comet would return in 1758. His prophecy was fulfilled for the comet was sighted in that year, though it did not appear at its brightest until 1759. Halley's comet appeared again in 1835 and in 1910, and will next be visible in 1986.

Further investigation showed that the comet had been observed at every appearance as far back as the year 240 B.C. It should have been visible in 315 B.C. and 391 B.C., but there are no actual records of its appearance in those years. However, the comet of 467 B.C., the first that was definitely recorded, was undoubtedly Halley's comet.

This comet figured prominently in ancient and medieval chronicles. It was usu-

ally discussed in connection with some calamity or other, which it was supposed to have foretold. It was seen in the year 66 of our era, a few years before the destruction of Jerusalem by the Roman Emperor Titus, and was described by the Jewish historian Josephus as the "Sword of God" over the doomed city. It appeared in 451 at the time of Attila's invasion of Western Europe and again in 1066 when William the Conqueror invaded England. In the eleventh-century Bayeux tapestry, commemorating William's conquest, Halley's comet occupies a prominent place.

When Halley demonstrated in 1682 that the comet which now bears his name was a celestial body moving around the sun in accordance with the law of gravitation, he dispelled once and for all the notion that comets have baleful significance as signs of the wrath of God. For this reason the year 1682, when Halley observed the comet later named after him, represents an important landmark in intellectual history.

Since Halley's time, the orbits of a

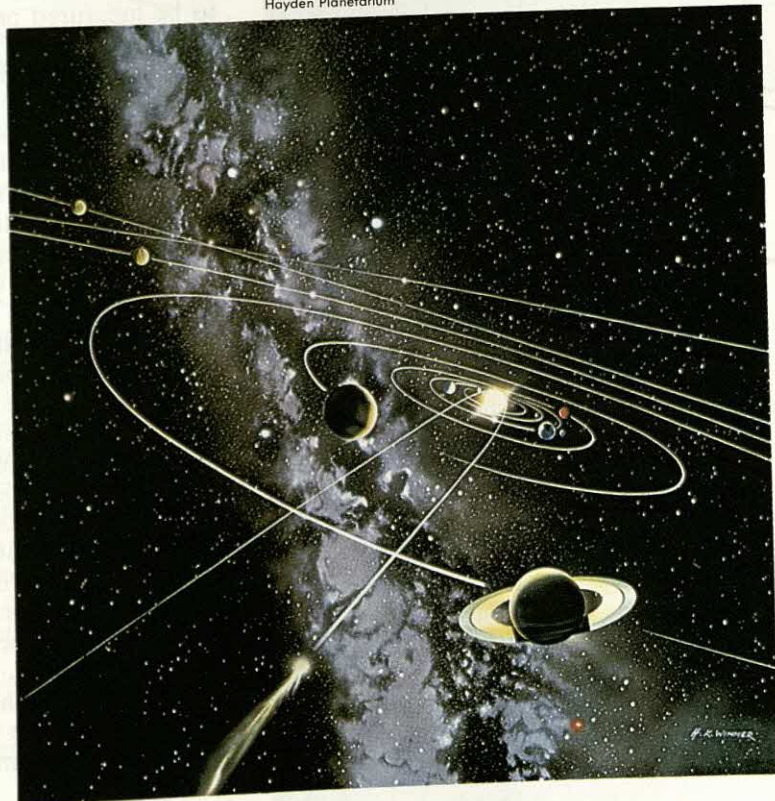
good many comets have been traced. To understand the nature of these orbits, it is necessary to point out that, according to the law of gravitation, comets must move around the sun in one of the conic sections. A conic section is a curve obtained when a plane cuts through a cone. There are only two such curves that are closed—the circle and the ellipse. A body moving in one of these curves ultimately returns to the place from which it started. As astronomers would say, the body has a definite period of revolution.

A comet cannot move in a circle. If it started by tracing a circular path, the attraction of planets in the solar system would soon distort its path into an ellipse. Thus, all periodic comets must move in ellipses.

The circle and the ellipse are not the only possible conic sections. The cone can be cut so as to produce open curves—parabolas and hyperbolas. A comet moving in one of these curves would travel around the sun and off into space, never to be seen again.

Painting by Helmut K. Wimmer, The American Museum of Natural History, Hayden Planetarium

A long cigar-shaped orbit of a comet is seen in this striking painting of the solar system as it would look from the vicinity of the planet Saturn.



At one point in its orbit, a comet will be nearest to the sun. This point is called *perihelion*. The opposite point in the orbit, farthest from the sun, is called *aphelion*. Halley's comet was at perihelion in 1910. In the year 1948 it was at aphelion.

NAMING COMETS

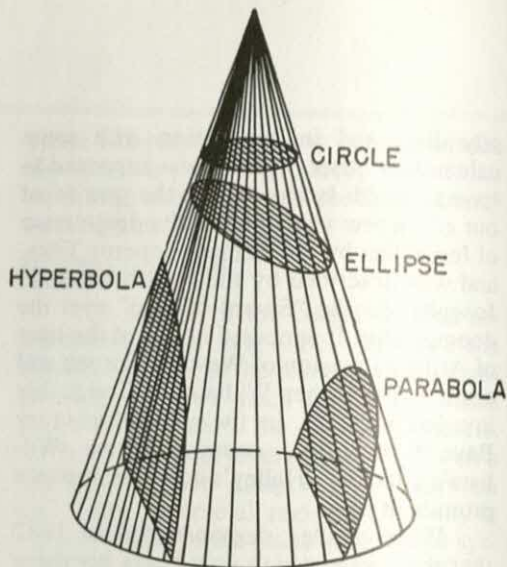
Astronomers must be able to identify comets, in order to calculate their orbits. Each comet, therefore, is distinguished by a special name. This contains the year when a particular appearance was first discovered, and a letter indicating the order of discovery among the comets of that year. This is a temporary name. Once the orbit has been computed and time of perihelion worked out, the same comet is designated by the year of perihelion passage, and a Roman numeral, indicating its order among the comets that have reached perihelion that year.

For example, Halley's comet in its last apparition was first called 1909c and then 1910 II. This means that it was the third comet discovered in 1909, and the second comet to pass perihelion in 1910. In the case of more recently sighted comets, the name of the discoverer is added in parentheses. Thus the bright comet temporarily designated as 1936a is now referred to as 1936 II (Peltier). Thus it was the first comet discovered in 1936; it was the second comet to pass perihelion in that year; the discoverer was Peltier.

News of a comet's discovery in the Eastern Hemisphere is immediately sent to the observatory in Copenhagen, Denmark, the international clearing house for information of this kind. When the Harvard Observatory receives word of the discovery from Copenhagen, it telegraphs the information to all the observatories in the Western Hemisphere. News of a discovery in the Western Hemisphere is sent to the Harvard Observatory, which then forwards the news to other observatories.

DETERMINING THE COMET'S ORBIT

After a comet is discovered, astronomers begin to measure the apparent position of the comet in reference to known



A conic section—a circle, an ellipse, a parabola, or an hyperbola—is obtained when a plane intersects a cone.

stars. Its position, measured at intervals of a few days, gives the astronomer enough information to calculate a preliminary orbit and to predict the comet's motion in the future. Some of these measurements may be in error, or the comet may be too diffuse to be measured precisely. As a result, the first calculation of the orbit is usually somewhat inaccurate. It is amended as further observations become available. When the comet has disappeared from view, all observations are collected, and then the definitive or final orbit is computed.

The main difficulty in determining an orbit is that at first we have no idea how distant the comet is. It is presumably "near" the earth (that is, within 160,000,000 kilometers), but it appears to be as remote as the stars. It is only when it moves against the background of the constellations that we can begin to calculate its actual distance from the earth. There is an added complication. The apparent motion of the comet among the constellations is due principally to its real motion around the sun, but it is also due in part to the motion of the earth during the interval between observations. Therefore the astronomer must disassociate, or separate, the motion of the earth from that of the comet.

At best, even with modern calculating machines, the computation of the orbit is

very tedious, and even the best computers make mistakes. Therefore two computers generally work independently and check their results from time to time. After the orbit has been computed, the *ephemeris* is set up. An ephemeris is a statement, in the form of a table, of the assigned places of a celestial body for regular intervals.

It may take a long time to work out the definitive orbit. Each observation must be checked, the best known position for each comparison star must be obtained, and the attraction (perturbation) by other members of the solar system must be allowed for. From all these computations various quantities are derived. They are called elements of the orbit. One of the elements is the time of perihelion. Another is the distance of perihelion from the sun, measured in astronomical units. An astronomical unit is a unit of measurement used by astronomers for great distances. One astronomical unit, abbreviated a.u., equals the distance from the earth to the sun, or approximately 149,000,000 kilometers.

SHORT AND LONG PERIODS

The time that elapses between two returns of a comet to perihelion is called its *period*. Some comets move in comparatively small ellipses, so that they come back to perihelion every few years; that is, they are short-period comets. Encke's comet has the shortest known period—about 3.3 years. It was first discovered in 1786 and has since been seen returning to perihelion more than forty times. None of the short-period comets are bright. Under the best conditions Encke's comet is barely visible to the naked eye.

If the orbits of certain short-period comets, about thirty in number, are plotted, it is seen that all of the aphelia, or positions farthest from the sun, fall near the orbit of Jupiter. Obviously this giant among the planets exerts an important influence upon these comets. It can be shown mathematically that they were deflected by Jupiter

from their original orbits and made to move around the sun in small ellipses. Therefore these short-period bodies are called Jupiter's family of comets.

Other planets have also deflected comets in much the same way and now have their own comet families. Saturn's family is few in number. It includes the remarkable Comet 1925 II (Schwassmann-Wachmann), which has a period of 16 years and an orbit that is almost a circle. It was the first comet to be observed in every part of its orbit, from perihelion to aphelion. This strange body is subject to remarkable changes in brightness, the cause of which is unknown. Sometimes it flares up and becomes more than 500 times as bright as it was a few

New Mexico State University Observatory



Comet Ikeya-Seki, visible in late 1965, was the first comet to have its temperature taken. Infrared techniques revealed that when 32 million kilometers from the sun, its temperature was 650° Celsius.

days previously. Unfortunately, it is always so far from both the earth and sun that even at its brightest a good telescope is necessary to see it at all.

Halley's comet belongs to Neptune's family, which, like Saturn's family, has only a few members. Some of these, including Halley's comet and Comet 1884 I (Pons-Brooks), are quite bright.

For some comets, periods of 100, 200, or 500 years have been calculated. The longer the period, the more uncertain it is. The trouble is that we can observe such comets only in the small part of their orbit near perihelion. In the case of comets with periods of more than a few thousand years, the orbit that can be observed shows exceedingly little curvature. This often makes it impossible to figure out the shape of the orbit or the length of the period. According to the Estonian astronomer E. Opik, some of the longer orbits may extend out to the nearest stars, more than four light-years away. A light-year is approximately 9,600,000,000,000 kilometers.

Most astronomers believe that all comets are periodic, although the periods may be many thousands or even millions of years. It is thought, for example, that Comet 1914 V (Delavan) has a period of 24,000,000 years. It is possible, however, that the orbits of some comets may form an open curve, whose ends can never meet. In such cases, the comet may have been traveling in an elliptical orbit when it entered the planetary region of the solar system, and may have been thrown into an orbit forming an open curve because of the gravitational attraction of the major planets.

COLLISION WITH A PLANET?

Since comets move in all directions in the solar system, it is conceivable that a comet might collide with a planet like the earth. We have already seen that no damage is done when the earth passes through a comet's tail, but what would happen if the earth were struck by the nucleus?

The chances of this taking place are very small. Yet in the earth's long history there must have been a few collisions with comets. In such an event there would cer-

tainly be a brilliant display of meteors as the comet's nucleus disintegrated upon coming in contact with the earth's atmosphere. The larger pieces of the comet's nucleus might cause considerable destruction if they fell into a populated region.

APPARENT BRIGHTNESS

As we have observed, the illumination of comets is due, directly or indirectly, to the sun. The nearer to the sun a comet is, therefore, the brighter it is. The distance of the comet from the earth will also naturally affect its apparent brightness. In 1910 Halley's comet appeared brightest, not at perihelion, but a month later, about the middle of May, when it came closest to our planet. The brightness of comets is measured by the scale of apparent magnitudes used to measure the apparent brightness of the stars.

One of the brightest comets ever seen in the skies was that of 1577. It was carefully observed by Tycho Brahe, who proved that it was not in our atmosphere. This demonstration dealt the death blow to the old theory that comets are vapors from the surface of the earth, ignited in the upper atmosphere of our planet.

A very bright comet appeared in 1680 and was studied in detail by Sir Isaac Newton in his famous *Principia*. Others appeared in 1744, 1811, 1843, 1858, 1861, and 1882. Comet 1858 VI (Donati), visible in the western sky in September and October 1858, was only moderately bright, but it was well situated for observation. People everywhere could see it.

Only a few bright comets have appeared in the 20th century, though there have been many easily visible to the naked eye. We have already mentioned Halley's comet, which appeared in 1910. Comet 1927 IX (Skjellerup) was very bright, but it was almost entirely ignored by the general public because it appeared to be so close to the sun. Another extremely bright comet was the one called Ikeya-Seki. Discovered by two amateur Japanese astronomers, K. Ikeya and T. Seki in September 1965, it created quite a sensation. After passing within 500,000 kilometers of the sun, the

nucleus of the comet broke into three pieces. These three pieces then sped back into space.

The brightness of a comet is difficult to predict. In the winter of 1973-74 Comet Kohoutek failed to appear as brightly as had been predicted, while in early 1976 Comet West surprised many, becoming one of the brightest comets observed since the beginning of the 20th century.

AMATEURS DISCOVER MANY

Many an amateur has been inspired by the appearance of a particularly spectacular comet to launch into a search for new comets. As a matter of fact, most new comets are discovered by amateurs. A professional astronomer using telescopes follows a rigid program of observation and for the most part observes individual stars or small parts of the sky. Once in a while a hitherto unknown comet will come within this field of view, but this does not happen often. The amateur who makes a deliberate search for a comet by scanning every quarter of the sky is more likely to make a discovery.

The amateur comet hunter should use a small telescope that has a wide field of view—four or five times the diameter of the moon—and should work out a definite system of observation. One excellent method is to sweep the sky from east to west and to change the position of the telescope with every sweep so that the new field partially overlaps the old one. The most promising areas for search are the western sky after nightfall and eastern sky before dawn.

The comet-hunting amateur will find many celestial objects that look like comets. Many of these, however, will prove to be nebulae and globular clusters. The only way to distinguish between such permanent objects in the sky and comets is to memorize the positions of the nebulae and globular clusters or to consult an atlas of the sky, such as Norton's. The real comet will betray its nature by changing its position among the stars within a few hours. Only occasionally will an observer come upon a comet with a tail about which there can be no doubt. In most cases the comet will be faint and diffuse, without any tail.

Only a fairly skillful amateur astronomer can search for new comets with any hope of success. The comet hunter must be able to operate a telescope efficiently, must know how to distinguish a comet from a nebula, and must be able to determine with what speed and in what direction the new comet is moving. Besides a certain amount of proficiency in astronomy, an amateur comet hunter must also have a great deal of persistence and must realize, too, that all efforts may be fruitless, for luck plays an important part in every discovery. But there are compensations. The searcher will gain a greater knowledge of the heavens while scanning the star-studded sky night after night for month after month. If the hunt is successful at last and a discovery confirmed, the hunter will have the immense satisfaction of knowing that a heavenly body will be known by his or her name for all time to come.

The champion discoverer of comets was Jean-Louis Pons, who had been a janitor at the Marseilles Observatory. Between the years 1802 and 1827 he discovered twenty-eight comets—more than one a year—with his homemade telescope. The American amateur comet hunters W. R. Brooks and E. E. Barnard attained such fame for their discoveries that they became professional astronomers. Brooks discovered twenty-five comets between 1883 and 1911, and Barnard in about the same period found twenty-two. Leslie Peltier won fame in the present century as a comet hunter. In a little observatory on his farm in Ohio, he added comet after comet to his "bag."

STILL PROBLEMS TO SOLVE

Much remains to be done in observing the spectra of comets and in determining their brightness, shape, and motions. The observation of comets from space—away from interference caused by the earth's atmosphere—now provides a new approach to the study of comets. In 1974 the study of comet Kohoutek by the then-orbiting U.S. Skylab crew added much information. Future observations from space will no doubt further advance our understanding of these unusual members of the solar system.



Harvard College Observatory

Asteroids are tiny planetary bodies that orbit the sun. Most are found between the orbits of Mars and Jupiter. In this three-hour exposed photograph, the trail, or part of the orbital path, of one asteroid is shown against a field of stars.

THE ASTEROIDS

In the region between the orbits of Mars and Jupiter, we find the orbits of a vast number of small celestial bodies. These bodies are known as asteroids, minor planets, or planetoids. About two thousand of these bodies have already been catalogued, and there may be over a hundred thousand. Some of them have orbits that swing beyond Jupiter, and some come inside the orbit of Mars.

The asteroids were discovered because astronomers were intrigued by an apparent flaw in the law they then used for estimating the relative distances of the planets from the sun. This law, formulated by the German astronomer Johann Elert Bode in 1772 and known as Bode's Law, was based on the fact that as the distance from the sun increases, the planet paths are more and more widely separated. Bode claimed that the increasing distances of the planets from the sun followed a more or less regular ratio. Today we realize that the law is based upon coincidence and in fact does not apply at all to the two outermost planets, Neptune and Pluto, which were not known in Bode's day. However, for a number of years after it was proposed, the law was accepted by astronomers.

The only difficulty Bode's Law seemed to present was the fact that the gap between Mars and Jupiter was too great to be explained by the law. Astronomers came to the conclusion that another planet, hitherto undiscovered, must lie within this belt. Toward the end of the eighteenth century, an association of astronomers was formed for the purpose of hunting for the missing planet. They realized that the object of their search must be small, or else it would not have escaped observation for such a long time.

DISCOVERY OF A "MISSING PLANET"

When Giuseppe Piazzi, an Italian astronomer, announced on January 1, 1801, that he had discovered a new heavenly body in the zone between the orbits of Mars and Jupiter, astronomers assumed that this was the planet for which they had been searching. Curiously enough, Piazzi came upon the supposed planet more or less by accident. He was engaged in making a catalogue of the fixed stars. He had developed a very exact method for mapping the sky by determining the relative positions of the stars within a given area on a number of successive occasions. If any

"star" moved in relation to its neighbors, it was obviously not a star at all but some other sort of heavenly body, such as a planet or a comet. The moving body would not be included, therefore, in Piazzi's catalogue of stars.

Piazzi had already mapped out more than 150 areas of the sky without incident. However, when he compared four successive observations of the constellation Taurus, he discovered that a certain small star within the constellation had changed its position from one observation to another. The Italian astronomer suspected that the "star" was really a comet. But when Bode heard of its movements, he decided that it was the planet for which so many astronomers had been searching.

THREE MORE "PLANETS"

Piazzi fell ill before he had time to make many observations of the new heavenly body, and it was lost to view for a time. Word of the discovery had come to Karl Friedrich Gauss, a young mathematician of Goettingen, Germany. Using a new method for determining planetary orbits, Gauss was able to calculate the path of the supposed planet from the few observations made by Piazzi. As a result, it was rediscovered in December 1801. It received the name of Ceres, after the Roman goddess of plant life. Not long after Ceres had been rediscovered in the heavens, astronomers discovered three other "planets"—Pallas (1802), Juno (1804), and Vesta (1807). It was realized by this time that the newly discovered heavenly bodies were too small to rank as full-fledged planets. Rather, they were minor planets. The names asteroids and planetoids have also been applied to them.

Ceres, Pallas, Juno, and Vesta are the Big Four among the asteroids. They are the only ones that show definite diameters and appear as disks when viewed through a telescope. Ceres is about 770 kilometers in diameter, Pallas 490 kilometers, Vesta 385 kilometers, and Juno 190 kilometers. Though Vesta is the third in size, it is the brightest, perhaps because its surface reflects sunlight much more effectively than

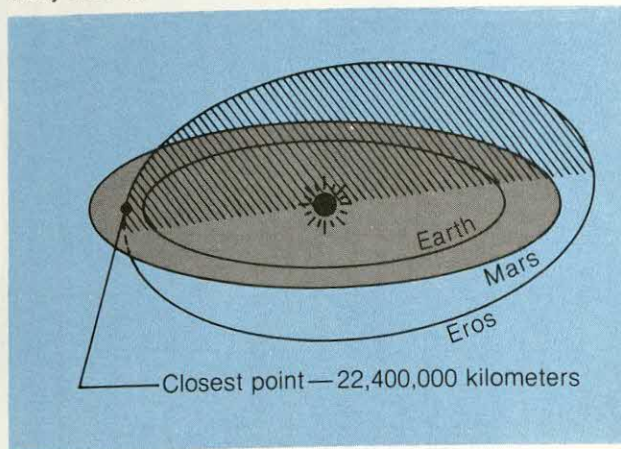
the surface of the other minor planets. Vesta is the only one that can be seen with the naked eye.

MANY MORE ON FILM

It was not until 1845 that the fifth asteroid, Astraea, was discovered. Astronomers were finding the search for the comparatively tiny asteroids painfully slow and laborious. But in 1891, a new asteroid-hunting technique was developed by Max Wolf, of Germany. He attached a camera to a telescope that was moved by clockwork in such a way that it continually pointed to the same fixed stars. A photographic plate was set in the camera and exposed for a certain period of time. When it was developed, the stars in the photograph appeared as white points. If there were any asteroid within the field of the telescope, it would appear as a short white line, because it would have moved in its orbit during the exposure of the plate. After adopting Wolf's technique, astronomers found so many new asteroids that it became difficult to keep track of them.

It was by Wolf's method of photographic observation that the small planetary body known as Eros was discovered in 1898 by the German astronomer G. Witt. Though a part of this asteroid's orbit lies outside that of the planet Mars, it is for the most part between the orbit of Mars and

The asteroid Eros swung close to earth in 1975, providing astronomers with a rare opportunity to study the tiny planet.



that of the earth. Eros, 24 kilometers in diameter, is an irregularly shaped body. It has a highly eccentric orbit and may approach to within 22,400,000 kilometers of the earth's path. By determining different positions of Eros during one of its approaches to the earth, astronomers have been able to calculate the astronomical unit—the distance of the earth from the sun—more precisely than by other methods.

VARIED ORBITS

The asteroids cover a very wide belt in space. Hidalgo, at its farthest distance from the sun, approaches the orbit of the planet Saturn. Icarus, discovered in 1948, moves in an orbit that passes beyond Mars and then closer to the sun than the planet Mercury itself—a distance of only 30,000,000 kilometers, which is about half that of Mercury from the sun. The asteroid Hermes, only $1\frac{1}{2}$ kilometers across, sometimes comes as close as 320,000 kilometers to the earth's orbit. This distance is smaller than that of the moon from the earth. In 1937, Hermes moved to within 800,000 kilometers of the earth itself. Evidently the orbits of the asteroids are highly varied and distorted compared to those of the planets. They may intersect each other and often lie at high angles to the planes of the earth's path and the paths of other planets around the sun. The asteroids have never been

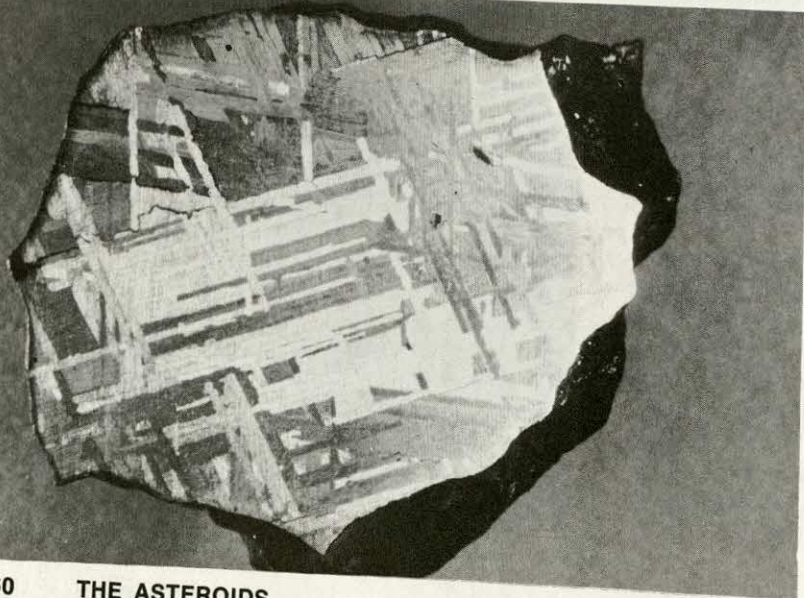
seen to collide with each other or with any of the planets, but this may have happened in the past.

The larger asteroids have measurable disks, but in the great majority of cases, the diameter can be determined only by indirect means. As we saw, the largest asteroid, Vesta, is only 770 kilometers across. The smallest ones may be only one or two kilometers in their longest dimensions. Probably, like Eros, they are not spheres but irregularly shaped solids.

ORIGIN OF THE ASTEROIDS

Astronomers have speculated on the origin of the asteroids. One theory, recently shown to be impossible, is that they are the remains of an exploded planet that had previously circled the sun between the orbits of Mars and Jupiter. The planet approached Jupiter too closely and was broken up by the gravitational pull of that giant planet. The fragments then collided, accounting for their varied orbits.

Another explanation of the origin of asteroids is that they are chunks of cosmic matter that somehow never finally united to form a planet, during the time that the solar system was coming into being. Here, too, the gravitational attraction of Jupiter would be the decisive factor. It would prevent the chunks from drawing together and ultimately forming a single body.



From the collection of
Ronald A. Orin,
Griffith Observatory.
Photo by James E. Klein.

Meteorites—perhaps stray asteroids that collide with earth—are analyzed to help answer questions concerning the nature and origin of asteroids. At left: an iron meteorite.

METEORITES AND METEORS

From time to time, even the most casual watcher of the night skies will observe a point of light, perhaps trailing a fleeting luminous train, in swift apparent motion against the background of the fixed stars. This luminous object may range in intensity from barely visible to a brightness rivaling that of the sun itself. The popular name for it is "shooting star." Actually, what the observer sees is not a far-off star in motion, but a phenomenon that takes place in the earth's atmosphere. It occurs when a swiftly moving body from outer space penetrates the atmosphere and becomes so heated from air resistance that it begins to glow.

The process by which it becomes luminous involves at least two distinct stages. First, there is an invisible stage, high up in the rarefied outer atmosphere. The scant molecules of air strike the surface of the invading mass like sleet beating violently against the nose of a speeding jet liner. Atoms are then ejected from the surface of the mass, which begins to melt and even to vaporize. Chiefly because of these vaporization effects, the outer layers of the invading body are transformed into a sort of miniature atmosphere which surrounds the body and substantially increases its size.

This ever-enlarging complex of solid, liquid, and gaseous matter now penetrates the atmospheric layers nearer the surface of the earth and enters a second stage. A vastly increasing number of collisions take place with the more and more numerous air molecules in the lower atmosphere. The energy of motion of the invading body is transformed into other forms of energy, in-

American Museum — Hayden Planetarium

A brilliant meteor seen in Czechoslovakia. The meteorite may disintegrate while travelling through the atmosphere or upon impact with the earth.





Department of Tourism, State of Arizona

Great Meteor Crater near Winslow, Arizona. This large depression has been recognized as a meteorite crater since the beginning of the 20th century.

cluding radiant energy. This may take the form of visible light. It is then that the swiftly moving complex becomes visible to the observer on earth as a "shooting star."

The object from outer space may be millions or even thousands of millions of years old before it strikes our atmosphere. But the luminous phenomenon it produces is fleeting. It lasts for only a fraction of a second in most cases, and only rarely exceeds a small fraction of a minute.

In this article, we shall apply the name *meteorite* to any natural solid body that originates in outer space and reaches the earth's surface without completely burning up. We shall use the term *meteor* to refer to those bodies that vaporize when they reach the earth's atmosphere. They appear as "shooting stars."

Meteorites provide us with specimens of objects populating outer space. From these specimens, we can determine directly the composition of matter existing beyond the earth and moon. We can also draw conclusions about the conditions under which this matter originated and evolved and about the time when it came into being.

SIZES OF METEORITES

According to recent estimates, 10,000 metric tons of meteorites reach the earth every day. Most of these bodies are exceedingly small—so tiny that they are not particularly affected by their passage through the atmosphere. Eventually, these particles drift down to the surface of the earth, almost or quite unaltered. Their longest dimension is not more than one-one hundredth of a centimeter. As these parti-

cles pass through the outer atmosphere, they can be collected by special devices installed on artificial satellites. These almost imperceptibly small specks are known as micrometeorites, or "tiny meteorites."

Next in size are the meteorites with a maximum dimension of about one centimeter. They make up the great majority of the objects that give rise to glowing light as they pass through the atmosphere.

Larger meteorites enter the atmosphere less frequently, but when they do they produce brighter streaks. And as the light effects become increasingly spectacular, sound effects start to occur. Meteorites weighing 4.5 kilograms or more are not completely destroyed by the rigors of atmospheric travel, and small but recognizable portions reach the surface of the earth. The light and sound effects produced by a meteorite of this size as it traverses the atmosphere range from the merely startling to the terrifying. The passage of large meteorites may produce an ultrabright fireball and violent shock waves underneath the line of flight. At the moment of impact, high-speed "meteoritic shrapnel" has been observed to cut off tree branches as smoothly as would a sharp axe. Meteorites have cleanly perforated sheets of hard, brittle material, such as roof slate, without cracking them. They have punched holes through layers of pond- and sea-ice and through the metal roofs of automobiles.

SPECTACULAR FALLS

In the late 1940's there were two spectacular meteorite falls. On the morning of February 12, 1947, a solid metallic meteor-



Royal Canadian Air Force

Deep Bay Crater in Saskatchewan, Canada, is about 12 kilometers in diameter. This large depression was probably formed when a huge mass collided with earth.

ite suddenly shattered and fell as a "shower of iron" in the Ussuri taiga, northeast of Vladivostok, in the Soviet Union. More than 120 craters—some large enough to hold a two-story house—were created. One of the recovered masses, weighing 1,745 kilograms, is the largest meteorite of witnessed fall so far recovered in the world. It is estimated that the Ussuri meteorite, known in the Soviet Union as the Sikhote-Alin meteorite, had a mass in excess of one hundred metric tons before it encountered the earth's atmosphere.

On the afternoon of February 18, 1948, a stony meteorite was seen to fall by tens of thousands of persons in Kansas, Nebraska, and adjoining states. The main mass of this fall, now known as the Furnas County stone, weighs more than one metric ton and is the largest stony meteorite so far recovered in the world. By the most conservative estimate, the total mass of the original meteorite before it struck the atmosphere must have exceeded 10 metric tons.

One other incident may have been the result of the fall of a meteorite many times larger than the Ussuri or Furnas County falls. On June 30, 1908 a fireball blazed across the sky in the Tunguska region of Siberia. Extremely violent airwaves (recorded all around the world) and earth tremors followed the meteorite's impact. In the central region of fall, the pressure effect blew down more than eighty million trees. The heat developed was so intense that

tree trunks and the carcasses of reindeer killed by the blast were charred.

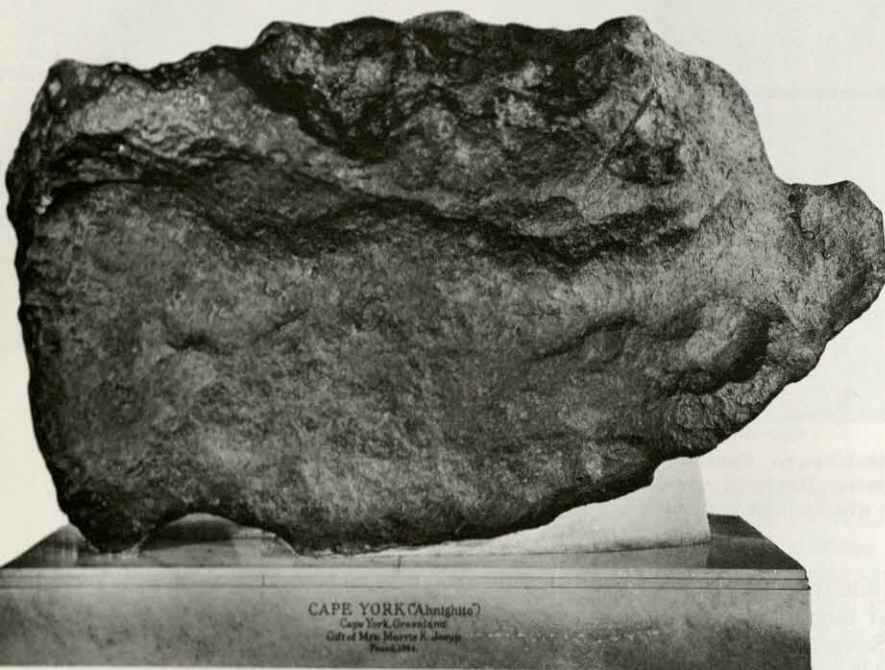
No meteorites have ever been recovered from this area. This almost incredible fact has led many scientists to question if the event was really caused by a meteorite. Numerous theories have been developed—some quite fanciful—but the mystery remains.

Only a few of all the meteorites so far recovered in the world measure up to the Ussuri and Furnas County meteorites that fell to earth within a year of each other. Such massive meteorites include a huge iron of perhaps 70 metric tons weight near Hoba, South West Africa; the 30 metric ton iron meteorite Ahnighito brought back from Greenland by the American explorer Robert Peary; and the Bacubirito iron of about 27 metric tons—a tremendous specimen from Mexico.

CALLING CARDS—CRATERS

Apparently, meteorites much larger than the original Ussuri mass would be almost completely destroyed upon striking the earth's surface. Experiments have shown that when a missile strikes a target at comparatively low speeds, the impact results in a narrow but fairly deep penetration funnel. At considerably higher impact speeds, the depth of the basin gouged out by the missile is not so great, but the increase in width more than offsets the decrease in depth. At even higher speeds the projectile and a considerable part of the target are volatilized. The resulting basin becomes a true explosion crater. Although it is relatively shallow, its volume is surprisingly large compared to the size of the projectile that created it.

Let us apply these findings to meteorites. If a meteorite were large enough so that it would be slowed down but little by the tenuous atmospheric shield, it would plunge into its target on the surface of the earth at a speed of from several kilometers to scores of kilometers per second. Both the meteorite and its target on earth would vaporize with frightful rapidity. The result would be an explosion of the most catastrophic violence. The explosion crater



This 30-metric-ton meteorite from Greenland is one of the largest ever found.

CAPE YORK (Ahngaitte)
Cape York, Greenland
Gift of Mrs. Maurice R. Joseph
Found, 1961

American Museum of Natural History

thus blasted out of the face of the earth would be vast in size, dwarfing even the biggest of the Ussuri craters. The name "megameteorites" ("giant meteorites") has been proposed for meteorites capable of producing such effects.

We would expect that "calling cards" in the form of vast craters would remain, providing evidence that a megameteorite had plunged into the earth. As a matter of fact, such evidence became available years ago when it was discovered that the vast craters at Canyon Diablo, Arizona, and Odessa, Texas, resulted from the explosive action of meteorites. The Canyon Diablo meteorite crater was identified as such in 1905. It has a circumference of nearly five kilometers and once was nearly 400 meters deep. The meteorite that blew out this vast hole in the earth had a mass estimated at between 2,300,000 and 7,000,000 metric tons. Geological evidence indicates that this huge meteorite came to earth more than 50,000 years ago. The meteorite explosion crater at Odessa, with a diameter of more than 150 meters was identified in 1929. It was probably produced over 200,000 years ago. In comparison with the huge craters at Canyon Diablo and Odessa, the Ussuri craters are quite insignificant.

NEAR MISSES

Even these megameteorites would, however, be dwarfed by the colossal meteorites known as minor planets, planetoids, or asteroids. Most asteroids orbit the sun between the paths of Mars and Jupiter, but some have orbits that bring them much closer to the earth. Already these minor planets have scored near misses with the earth. In 1936, the minor planet Adonis passed the earth at a distance of only 1,500,000 kilometers. The next year, another minor planet, Hermes, came still closer to the earth, approaching to within 800,000 kilometers.

It seems probable that some of the very numerous minor planets weighing no more than a few hundred million metric tons or a few thousand million metric tons have actually scored direct hits on earth during the long geologic past.

CLASSIFICATION OF METEORITES

The regions around meteorite craters are rich sources of meteoritic materials. Many metric tons of meteorites have been removed from the large crater areas. Lumping together all kinds of meteorites, we find that the following ten elements make up

more than 99 per cent (by weight) of meteoritic material:

| <i>Element</i> | <i>Percentage by Weight</i> |
|----------------|-----------------------------|
| Oxygen | 34.6 |
| Iron | 25.6 |
| Silicon | 17.8 |
| Magnesium | 13.9 |
| Sulfur | 2.0 |
| Calcium | 1.6 |
| Nickel | 1.4 |
| Aluminum | 1.4 |
| Sodium | 0.7 |
| Phosphorus | 0.16 |
| | 99.16 |

Other rare minerals and some naturally created isotopes have also been found. All meteorites universally recognized as such have generally been placed in one or the other of the three following groups: (1) irons, or siderites; (2) stones, or aerolites; and (3) stony irons, or siderolites.

Irons, or *siderites*, are about 91 per cent iron and 8.5 per cent nickel, on the average. They also contain cobalt, phosphorus, and minute quantities of other elements.

Stones, or *aerolites*, contain, on the average, 41 per cent oxygen, 21 per cent silicon, 15.5 per cent iron, 14.3 per cent magnesium, and smaller percentages of other elements. The aerolites are much less dense than the irons and are much more fragile. Consequently, they disintegrate more or less completely on their way through the resisting atmosphere and fall in widely scattered showers of small masses, only a few of which are recovered. Aerolite falls have been observed about 10 times more frequently than iron meteorite falls.

Stony irons, or *siderolites*, show characteristics of both the iron and the stony meteorites. Stony irons are much less common in collections than either siderites or aerolites and less is known about them.

Some scientists believe that we should set up a fourth division, that of the wholly glassy objects called *tektites*. Tektites resemble weathered pebbles of glassy rock. They are found principally along three great circles, or paths around the earth's sphere.

The meteoritic origin of tektites is not firmly established, but some scientists believe that tektites are the products of collisions between the earth and orbiting meteorites.

CARRIERS OF LIFE?

There was once wide acceptance of the theory that meteorites carry living spores and thus serve to propagate life in different parts of the universe. Today this theory is no longer held by experts. However, some authorities believe that meteorites may contain nonliving evidence of organisms that once lived far beyond the earth.

Organic, or carbon-containing, compounds have been found in meteorites, often in highly organized complexes that resemble cells. Although some scientists believe that this is strong evidence of other life in the universe, these complexes can be explained by other hypotheses.

Over a long period of time they could have formed in space, where the component elements—carbon, hydrogen, nitrogen, and oxygen—are known to exist. They could also result from contamination by earthly organisms. These contaminants—the cell-like complexes—could have entered the meteorite as it passed through our atmosphere, or as it rested on the ground.

AGES OF METEORITES

Methods comparable to those used in calculating the ages of selected earth rocks have been used to determine the age of various meteorites. Some of the results indicate the time that has elapsed since the meteorites solidified. In other cases, the ages deduced are exposure ages: that is, measures of the length of time the test-specimens were exposed to cosmic radiation in space. Some meteorites are reported to be less than 75,000,000 years old. The ages of others may be billions of years old. According to one estimate, the meteorite from Mount Ayliff, South Africa, is nearly 7,000,000,000 years older than the earth or the sun.

METEOR SHOWERS

Unlike unpredictable meteorite falls, some groups of meteors appear at more or

SOME WELL-KNOWN METEOR SHOWERS

| NAME OF METEOR SHOWER | DATE OF MAXIMUM SHOWER ACTIVITY | AVERAGE VISUAL HOURLY RATE (AT MAXIMUM) | REMARKS |
|-------------------------------|---------------------------------|---|--|
| Quadrantids | January 1-3 | 25-75 | Annual shower. Medium-speed meteors. Parent comet: perhaps Kozik-Peltier (1939 I). |
| Lyrids | April 21 | 5-60 | Annual shower. Swiftly moving streaks. Parent comet: Thatcher's (1861 I). |
| Eta-Aquarids | May 4-6 | 10-35 | Annual shower. Well visible in low latitudes. Parent comet: probably Halley's. |
| Delta-Aquarids | July 28 | 2-15 | Annual shower. Meteors slow, with long paths. |
| Arietids | June 8 | -- | Daylight meteor stream. Apparently connected with Delta-Aquarids and produced by same extended meteor stream. Hourly rate based on radio-echo observations: 10-60. |
| Perseids | August 10-17 | 35-70 | Annual shower. Swiftly moving meteors. Parent comet: Swift-Tuttle (1862 III). |
| Giacobinids (or Nu-Draconids) | October 9 | Very low to 6,000 or more | Periodic shower. Meteors often characterized by bright terminal bursts. Parent comet: Giacobini-Zinner. 6½-year period. |
| Orionids | October 20-23 | 5-20 | Annual shower. Swiftly moving meteors. Parent comet: very probably Halley's. |
| Taurids | November 3-10 | 5-20 | Annual shower. Slow meteors; some fireballs. Parent comet: Encke's. |
| Beta-Taurids | June 30-July 2 | -- | Daylight meteor stream. Identified as summer daytime return of November Taurid stream. Parent comet: Encke's. Hourly rate based on radio-echo observations: 10-25. |
| Leonids | November 16-17 | Very low to 30,000 or more | Periodic shower. Very swiftly moving meteors. Parent comet: Tempel-Tuttle (1866 I). 33-year period. |
| Bielids (Andromedids) | November 27-December 4 | Very low to 100-400 | Erratic shower. Parent comet: Biela's. |
| Geminids | December 12-13 | 40-60 | Annual shower. Medium-speed meteors. |
| Ursids | December 22 | 20-100 | New erratic annual shower. Parent comet: probably Tuttle's (1939k). |

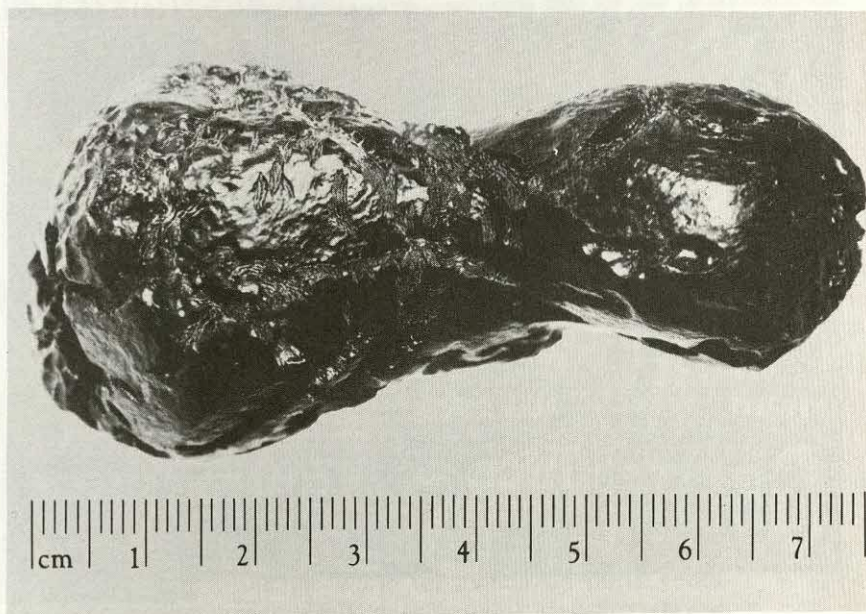
less regularly spaced intervals. The resulting displays are called *meteor showers*. The frequency of meteoric appearance in a shower at the period of maximum activity varies greatly. There may be as few as five or even less an hour, and the maximum usually does not exceed 75 or so. Sometimes, however, the rate of appearance is very much greater: thousands of meteors may be observed in the course of an hour. It is estimated, for example, that some 35,000 or so meteors fell in an hour during the Leonid meteor shower of November 1833. This spectacular display, which constituted a veritable meteoric cloudburst, terrified al-

most everyone who saw it. Other spectacular meteor showers included the Giacobinid showers in October 1933 and 1946 and the Leonid shower in the early morning of November 17, 1966.

In some cases, meteor showers are named after comets. Thus the Giacobinids are named after the Giacobini-Zinner comet; the Bielids, after Biela's comet. Generally, however, the constellation in which the radiant, or visible path, of a meteor shower is situated determines the name given to the shower.

It has often been found that the orbits of meteor showers correspond to those of

Tektites may have formed from rocks which were melted when struck by large meteorites.



Betty Jane Harbour

known comets, and authorities generally consider, therefore, that the showers represent debris from existing comets.

"NOW YOU SEE THEM . . ."

The study of meteors is important. Fortunately, the ability of the human eye to detect even rapidly moving luminous objects has permitted effective visual observation of meteors from the earliest times. Properly organized, such visual work allows fairly accurate determination of the atmospheric trajectories of meteors. On occasion, it has pinpointed areas in which successful searches for meteorite fragments were later made.

With the gradual improvement of photographic techniques in the early 20th century, the photographic recording of meteors finally began to provide worthwhile results. Later, radar techniques were adopted. Radio signals are sent out into space and are reflected back to earth from the ionized gases formed as the meteor passes through the atmosphere. They are then recorded in a receiver. Thus it becomes possible to detect meteors whether they appear at night or during the daytime.

Determination of the real path that meteors follow through the earth's atmo-

sphere enables us to find their velocity. Then we can determine the nature of the orbit in which the parent body moves.

Through the study of the real paths meteors follow through the earth's atmosphere, we can estimate the speeds with which these objects move around the sun before they crash into the atmosphere and are observed as meteors. Once we know the heliocentric, or sun-centered, velocity of one of these bodies, we can determine whether or not it is a member of the solar system. If its velocity at the earth's distance from the sun is less than 42 kilometers a second (approximately), then the object is a member of the solar system. It moves about the sun in a closed elliptical orbit. If, on the contrary, the heliocentric velocity of the body exceeds 42 kilometers a second, its orbit about the sun is an open curve, and the body is not a member of the solar system. If it had not collided with the earth, it would have been only a transient visitor from interstellar space. Twenty four percent of all meteors sighted are not members of our solar system.

How did both solar system and non-solar system meteorites originate? Final solutions to this and other problems of meteorite study are still being studied.



Modern calendars compete for popularity with colorful pictures that feature famous characters.

THE CALENDAR

by Elisabeth Achelis

If we asked the average person to name the things that are most essential in the daily round of his activities, the chances are that he would not include the calendar among them. Yet in every civilized age, ancient and modern, the calendar has been indispensable.

The calendar of today is the product of a great many centuries of patient study and of constant trial and error. When people first looked to heavenly bodies for a way to measure time, they observed that the sun seemed to make a constantly repeated journey in the heavens, always returning to the same place after many days. (Actually, of course, it is the earth that makes a yearly revolution around the sun.)

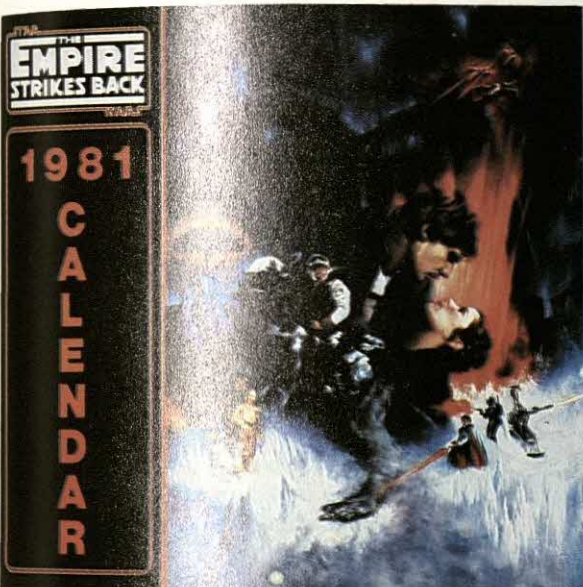
They observed that the moon also went through a cycle.

Most of the earliest calendars were based on moon cycles. These calendars were made to fit as best they could within the larger framework of the sun cycle. The

year, in these calendars, generally consisted of twelve moon cycles, or months. Since twelve moon cycles are not quite equal to a solar year, an extra month—called an intercalary, or inserted, month—was added from time to time. A number of ancient peoples, including the Babylonians, Hebrews, Greeks, and Romans, adopted this method of computation.

FIRST SOLAR CALENDAR

The Egyptians were the first to base their calendar on the sun cycle and to make the month a purely arbitrary unit, not corresponding to the actual lunar cycle. They worked out a year of 360 days, with 12 months of 30 days each. Since, according to their reckoning, it took 365 days for the sun to complete its journey in the heavens, they added 5 days to the end of the 360-day year. These added days were "feast days." The Egyptian priests were entrusted with the task of arranging for them.



© Lucasfilm Ltd. 1980 Used by permission of Ballantine Books, a Division of Random House, Inc.

This Egyptian 365-day calendar was adopted in the year 4236 B.C., according to the reckoning of the U.S. archaeologist James Henry Breasted. According to Breasted, it was the “earliest known and practically convenient calendar of 365 days.” As for the year 4236 B.C., it marked “not only the earliest fixed date in history but also the earliest date in the intellectual history of mankind.”

In the course of the centuries that followed, it was discovered that the year really consisted of $365\frac{1}{4}$ days. This additional quarter of a day was causing a gradual shift of the seasons as recorded in the calendar. In 238 B.C. the pharaoh Ptolemy III, also known to history as Euergetes I, tried to correct this error in calculation by adding another day to the calendar every four years. It was to be a religious holiday, but unfortunately, the priests were unwilling to accept the extra day. As a result the Egyptian calendar continued to be defective as a measure of the seasons.

THE MAYAN CALENDAR

Another seasonal sun calendar that was used in antiquity was that of the Mayas of Mexico. It probably goes back to the year 580 B.C. It was the first seasonal and

agricultural calendar in America.

The Mayan calendar was arranged differently from that of the Egyptians. Their solar year, called a *tun*, had 18 months of 20 days. It had a five-day unlucky period at its end to make 365 days. Each month had its own name, and the days were numbered from 0 to 19.

Dovetailed with the Mayan sun calendar was a religious year, sometimes called a *tzolkin*. The *tzolkin* contained 13 months of 20 days. Each day had a name that was combined with the numbers 1 to 13 to count out the 260 days of the *tzolkin*.

THE JULIAN CALENDAR

The Egyptian sun calendar was most carefully guarded by rulers and priests and consequently remained unknown to the outside world for more than thirty centuries. Only during Julius Caesar's stay in Egypt did he learn of this calendar, which was immensely superior in every respect to the one used in Rome.

The ancient Romans had a moon calendar. It was complicated and most confusing. There were 12 months; a thirteenth month, called Mercedonius, was occasionally inserted in a haphazard way. The 12 months of the Roman year consisted of 7 months of 29 days each, 4 months of 31 days each, and one month, Februarius (February) with 28 days, making a year of 355 days. The names of the 12 months of the Roman year were as follows:

| <i>Name of month</i> | <i>Origin of name</i> |
|----------------------|---|
| Martius | Month of Mars |
| Aprilis | “Opening” month, when the earth opens to produce new fruits |
| Maius | Month of the great god Jupiter |
| Junius | Month of the Junii, a Roman clan |
| Quintilis | Fifth month |
| Sextilis | Sixth month |
| September | Seventh month |
| October | Eighth month |
| November | Ninth month |
| December | Tenth month |



One of the most extraordinary works of Gothic art is the illuminated calendar *Très Riches Heures*, made about 1416.

Art Reference Bureau

| | |
|------------|---|
| Januarius | Month of the god Janus |
| Februarius | Month of the Februa, a purification feast |

In 153 B.C., January was designated as the first month of the year instead of Martius.

The Romans used a complicated system of reckoning within the month. There were three more-or-less fixed dates—the calends, the ides, and the nones. The calends always fell on the 1st of the month. The ides came on the 15th in Martius,

Maius, Sextilis, and October and on the 13th in other months. The nones always came on the 8th day before the ides. In designating a particular day of the month, Romans always reckoned backward from the calends, the ides, or the nones, as the case might be.

The calendar was entrusted to a council of priests—the College of Pontiffs, presided over by a *pontifex maximus*. The pontiffs were state officials charged with the regulation of certain religious matters, including the fixing of dates for ceremonies and feast days.

Caesar was elected *pontifex maximus* in 63 B.C., but it was not until 47 B.C. that he took the first steps to reform the calendar. Following the suggestions of the famous Greek astronomer Sosigenes, Caesar adopted the solar year for the Roman calendar. He gave it 365 days, plus a quarter-day of six hours. Quarter-days were withheld until a full day had accumulated. The day was then added to the common year as a leap-year day. This happened once every four years.

The year B.C. bridged the old and the new calendar. The following year, 45 B.C., was actually the first one using the reformed calendar. Caesar retained the complicated system of calends, nones, and ides within the months. January continued to be the first month of the year. The Roman Senate changed the name of the month Quintilis to Julius (our July) in honor of Caesar. The new calendar was known as the Julian calendar. Later, the Roman Senate changed the name of the month Sextilis to Augustus (August) to honor the Emperor Augustus.

THE SEVEN-DAY WEEK

In 321 A.D. the Emperor Constantine issued an edict introducing the seven-day

week in the calendar, doing away once and for all with the system of calends, ides, and nones. Constantine established Sunday as the first day of the week and set it aside as the Christian day of worship.

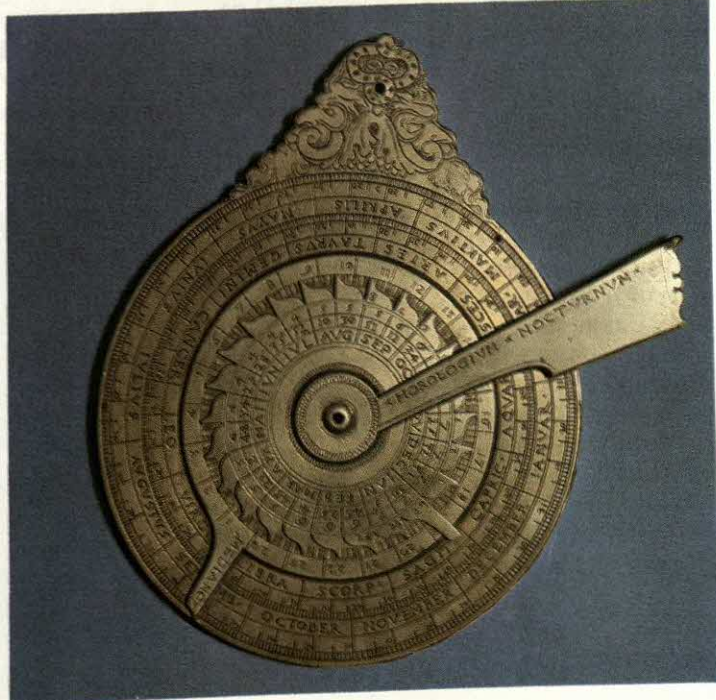
Although the introduction of the week greatly simplified matters, it brought about a serious defect in the calendar, which is still present today. Both the Egyptian and Julian calendars had been stabilized. That is, every year had been like every other year. Through Constantine's reform the Julian calendar became a shifting one. Now that there were 52 seven-day weeks, totaling 364 days, there was always one day left over in ordinary years and 2 days in leap years. This meant that in successive years, the Julian calendar began on different days of the week.

THE GREGORIAN CALENDAR

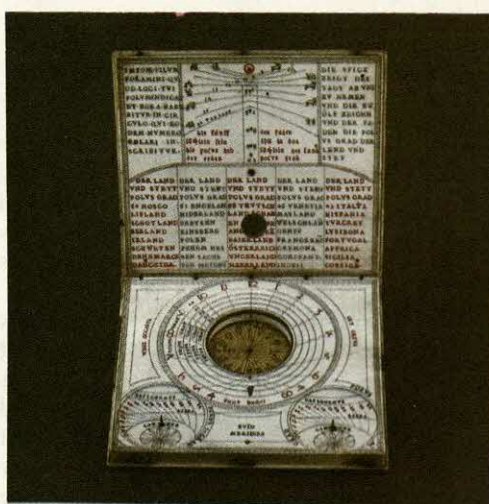
The true length of the solar year is a trifle less than $365\frac{1}{4}$ days. It is 365.242199 days, or 365 days, 5 hours, 48 minutes, and 46 seconds, to be exact. Therefore the Julian calendar was too long by about 11 minutes. After a number of centuries the error amounted to several days.

In the year 1582 another momentous calendar reform took place. Pope Gregory

A 16th century Florentine night dial in brass. It has a sun dial with a quadrant on the back.



G. Tomsich/Photo Researchers



G. Tomsich/Photo Researchers

This beautiful Italian Renaissance calendar has both a sun- and moon-based system to measure time.

XIII determined to adjust the calendar to the seasons. For this purpose he called upon the services of the mathematician Christopher Clavius and the astronomer-physician Aloysius Lilius. They found that the error caused by the excessive length of the Julian calendar amounted to ten days. To set the year aright, they canceled ten days from the Julian calendar, so that October 4, 1582, was followed by October 15. This was how the month looked on the calendar:

| 1582 | OCTOBER | | | | | | 1582 |
|------|---------|-----|-----|------|-----|-----|------|
| SUN | MON | TUE | WED | THUR | FRI | SAT | |
| | 1 | 2 | 3 | 4 | 15 | 16 | |
| 17 | 18 | 19 | 20 | 21 | 22 | 23 | |
| 24 | 25 | 26 | 27 | 28 | 29 | 30 | |
| 31 | | | | | | | |

Naturally this loss of ten days in the month of October created a certain amount of confusion. To avoid confusion, dates prior to October 15, 1582, were often given thereafter (and are still often given) as O.S. (old style) and dates after as N.S. (new style). If neither O.S. nor N.S. is given after a date, it's assumed to be N.S.

To avoid further error in the calendar, the leap-year rule was changed. In the case of centurial years (those ending in "00"), only the ones that were divisible by 400 were to be leap years. Noncenturial leap years continued to receive an extra day. No attempt was made to equalize the lengths of the months or to stabilize the

calendar. This Gregorian calendar is the one that we use today.

All Roman Catholic countries adopted the Gregorian reform, but other groups in Christendom were slow in accepting it. The English did not adopt the Gregorian calendar until 1752. France, like the other Catholic countries of Europe, had adopted the Gregorian calendar in 1582, but for a period from 1792 it was replaced by the "Revolutionary Calendar." In line with other anti-religious developments of the time, the days and months of this calendar were given symbolic names, such as "Brumaire" ("month of fog"), denoting the "natural" order of things. In 1806 Napoleon restored the Gregorian calendar as a gesture of reconciliation toward the Church.

Japan adopted the Gregorian calendar in 1873, China in 1912, Greece in 1924, and Turkey in 1927. Russia began to use the calendar in 1918, replaced it by another calendar when the Bolsheviks took over the country, and returned to the Gregorian calendar in 1940.

OTHER CALENDARS IN USE

The Gregorian calendar is not the only one used at the present time. For religious purposes Jews employ the Hebrew calendar, which begins with the year of creation, set at 3,760 years before the beginning of the Christian Era. This calendar is based on

Detail of an Aztec ritual calendar used by the Aztec priests to plan ceremonies and foretell favorable and unfavorable days.

Bibliothèque de l'Assemblée nationale, Paris



Right: an 18th century Dakota Sioux calendar shown on a buffalo skin.



Tom McHugh/Photo Researchers

the cycles of the moon. There are 12 months, which are alternately 29 and 30 days in length. An extra month of 29 days is intercalated 7 times in every cycle of 19 years. Whenever this is done, one of the 29-day months receives an extra day. The year begins in the autumn.

Another important calendar is the Islamic, or Moslem, calendar. It also is based on the cycles of the moon. There are 354 days and 12 months, half of which have 29 days and the other half 30. Thirty years form a cycle; 11 times in every cycle an extra day is added at the end of the year. The Moslem calendar begins with the first day of the year of the Hegira—that is, the journey of Mohammed to Medina. This date corresponds to July 15, 622, of the Christian Era.

Although the Gregorian calendar is China's official calendar, the Chinese New Year is still calculated by the ancient Chinese lunar calendar. The months of this lunar calendar are popularly known by the names of the 12 animals of the Chinese zodiac: rat, ox, tiger, hare, dragon, serpent, horse, sheep, monkey, rooster, dog, and boar.

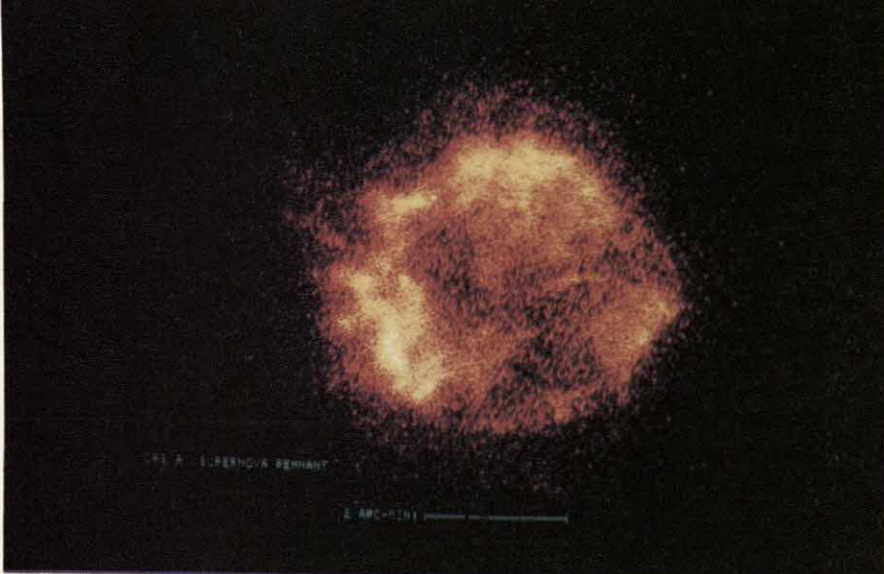
MODERN CALENDAR REFORM

The Gregorian calendar has served people well for almost four centuries. Yet some thoughtful people have tried to bring about reforms that would restore stability

to the calendar within the framework of the seasonal year.

In 1834, Abbé Marco Mastrofini put forward a plan in which every year would be the same and the lost stability of the calendar would be restored. In his calendar, there were 364 days in the year—a number easily divisible in various ways. The 365th day and the 366th in leap years were inserted as extra days within the year. Each year would begin on Sunday, January 1. The abbé's idea was so simple that most modern calendar reformers have made it the basis of their own proposals.

Calendar reform lagged until the League of Nations took up the question in 1923. One proposal, the World Calendar, based on the easily divisible number 12, once seemed particularly promising. In this calendar, each equal quarter year of 91 days, or 13 weeks or 3 months, corresponds to a season period. Every year in this calendar is like every other year. The first of every year, for example, falls on a Sunday; Christmas, December 25, falls on a Monday. To provide the necessary 365th day, a day—known as *Worldsday*—is inserted after December 30 and before January 1. The 366th day in leap years is inserted between June 30 and July 1. The Gregorian four-hundred centurial leap-year rule is retained. The idea was not, however, adopted and it has met with scant success in the years since.



X radiation glows in this spectacular image of Cassiopeia A, a supernova remnant. It is an expanding shell of gas and debris that was left when a star exploded.

© 1980 Smithsonian Institution

X-RAY ASTRONOMY

by F. R. Harnden, Jr.

Most of us have seen an X-ray picture—perhaps one showing a broken bone. Many of us also think it would be exciting to have X-ray vision and be able to see inside all sorts of things. A new field of astronomy—X-ray astronomy—has inspired that kind of excitement. It doesn't look inside things. But it is providing a fascinating new perspective on objects in the sky.

X-ray astronomy examines the skies in a new way. It studies something that is invisible to the eye—X rays. This new way of "looking" at the skies has brought into focus a violent and turbulent universe. It is a realm inhabited by hot young stars, by exploding galaxies, and perhaps even by black holes.

The recent development of this new branch of astronomy has been made possible by the space age. It is now possible, through the use of rockets and satellites, to observe things in space that cannot be studied from the earth's surface.

WHAT ARE X RAYS?

X rays are like light rays. Both are forms of radiation that travel in straight

lines at a speed of 300,000 kilometers per second. This radiation acts like waves in some respects. X rays, then, can be described by giving their wavelengths. The wavelength tells how far apart successive waves are. X rays have wavelengths 4,500 times shorter than those of visible light. X rays can be produced by actions among atoms and by other processes.

DETECTING X RAYS

X rays cannot travel through air as well as light rays do. This is because X rays ionize the atoms that make up the air. (An ionized atom is one that has lost or gained electrons.) When this ionization occurs the X rays are stopped. This means that X rays coming toward earth from space don't travel very far into the atmosphere. The blanket of air surrounding the earth blocks them. It also protects us from the radiation, which would be very harmful to life. Indeed, because X rays from space never reach the ground, they can only be observed from very high altitudes, above all or at least most of the atmosphere.

In the late 1940's scientists fitted

rockets with crude X-ray detectors and shot them straight up. During the brief moments when the rockets were at their highest points, the special instruments detected X rays coming from the sun.

For nearly 15 years the sun was the only object bright enough to be observed. But by 1962 better X-ray detectors and improved rockets had been developed. In 1962 a rocket detected a strange X ray-emitting object in the constellation Scorpius. For the first time, X rays from outside the solar system had been discovered, and X-ray astronomy was born.

The decade following the discovery of X rays from space saw tremendous activity. Rockets shot up to map the X-ray sky. So did high-altitude unpiloted research balloons. Many X-ray sources in our Milky Way galaxy and a few outside our galaxy were discovered.

It was an exciting time for X-ray astronomers. Nearly every observation meant the discovery of something new. This new-fangled branch of astronomy needed weeks and months of painstaking effort to build delicate instruments. These would then be rocketed into the air, only to come crashing back to earth a few minutes later. But if the experiment worked properly, the few minutes of observation were well worth the months of work.

Finally, in 1970, the first satellite designed to study X-ray emissions was launched. Named *Uhuru*, this satellite carried a simple scanning device that could tell bright from dark X-ray areas. *Uhuru*, however, could not really "see." The early instruments used in X-ray astronomy were like the light meters used by photographers. They could tell whether a patch of the sky was dim or bright in X rays, but they could not take pictures.

Then scientists learned how to convert X rays to visible light. They learned how to make X-ray pictures. With this advance X-ray astronomy took a giant leap forward.

MAKING AN X-RAY PICTURE

The first step in making X-ray pictures is to focus the X rays from space into an image. The lenses and mirrors used in an

optical telescope to focus light rays from space cannot be used to focus X rays. The X rays are simply absorbed by or sink into the lens or mirror.

A clever "trick," however, can be used to make a mirror focus X rays. Much as a stone can be made to skip across the surface of a pond, so can X rays be made to bounce off a mirror if they hit it at just the right angle. This system requires two very smooth mirrors that have been made in precise shapes. It focuses X rays of a certain wavelength very well.

The second step in making X-ray pictures is to turn the radiation into an electronic signal. One good way of doing this is to use a device known as a *proportional counter*. This device counts individual X rays. Each count produces a signal. The strength of the signal is proportional to the wavelength of the X ray that produced it. The signals are then fed to electronic devices that change them into a type of numeric code.

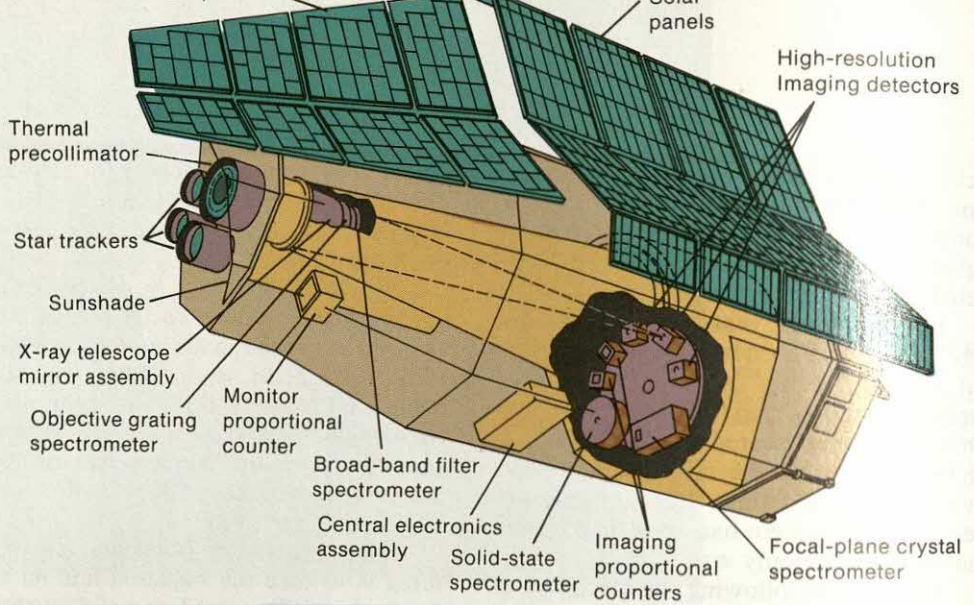
The final step in making X rays visible is to decode these signals. A computer connected to a color television system provides a good decoding system.

The X-ray image begins as a pattern of X rays in the telescope. It ends up as the same pattern translated into visible light by the television screen. The computer can also be programmed to change the colors used. Doing this enhances the contrast or highlights certain features of the image.

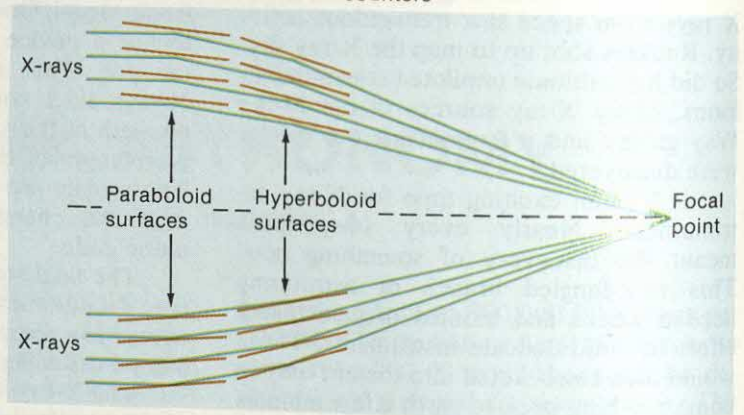
THE EINSTEIN OBSERVATORY

The first telescope capable of taking X-ray pictures of the stars was launched in 1978. It was part of the U.S. National Aeronautics and Space Administration (NASA) High Energy Astronomy Observatory (HEAO) program. The telescope and its satellite were the second in the program—HEAO-2—and were named the Einstein X-Ray Observatory.

The Einstein telescope consisted of an X-ray mirror like the one described above, two X-ray cameras, and two devices to analyze the component wavelengths in an X-ray picture. The telescope and spacecraft were controlled during nearly two and



The Einstein Observatory (above) is one of NASA's most advanced scientific spacecraft. Its spectrometers provide a wealth of information on the energy of the X-ray sky. The diagram at right shows how an X-ray telescope works. When X rays are reflected from specially-angled surfaces, the rays come to a focus.



one-half years of operation by on-board computers and by radio commands from a ground-based control center. From its orbit some 500 kilometers above the surface of the earth, the observatory each day recorded X-ray emissions from ten or twelve regions of the sky. It detected many thousands of X-ray sources, some 1,000 times fainter than any previously observed.

X-RAY EMITTERS

The first questions asked by X-ray astronomers were: Is there anything out there besides the sun to get excited about? If so, what is it? Where is it? How does it make X rays? Then astronomers began to realize first that there were dozens and then that there were hundreds and even thousands of X-ray objects. It became obvious that there must be many different answers to how X rays are made and what it is that makes them.

X-ray emitters, they found, include normal stars, young and old; neutron stars and black holes; galaxies; quasars; and clusters of galaxies.

Normal Stars, Young and Old. For decades astronomers have known that the sun is a rather ordinary star. There are many stars in the sky with properties like our sun. But the fact that X rays were first detected from the sun did not automatically mean that we would find X rays coming from similar middle-aged stars. The Einstein Observatory took pictures of hundreds of normal stars. It found that these stars give off far more X rays than had been expected. It found, in fact, that almost all stars—from hot young supergiants to normal stars like the sun to tiny dwarf stars—give off X rays.

Astronomers believe that many of the differences among stars are due to their ages. Stars are born out of vast clouds of

gas. Once they have formed, normal stars like the sun may spend billions of years shining rather serenely.

The births and deaths of many stars are, however, anything but serene. The more massive a star is when born, the more turbulent its early evolution can be. Huge stars, perhaps as much as 100 times as heavy as the sun, become very hot as they form out of gigantic clouds of gas and dust. This intense heat makes them strong emitters of X rays. Many pictures of such young stars were obtained with the Einstein telescope.

A star reaches the end of its productive life when the nuclear fuel in its interior is exhausted. Some stars simply fade away. They become ever dimmer as they grow cooler and cooler. Others go out with a bang. They meet their energy crisis with unimaginable violent explosions. The explosions can blow apart an entire star, spreading the "ashes" of the nuclear fires far into space.

It is this violent end of a star's normal life that can be studied particularly well by X-ray astronomy. Optical astronomers may witness the blast, referring to such an event as a *supernova*. After a few months, though, the optical star fades out. It is then that the blast wave from the explosion begins to heat up the space around the former star. This causes X rays to be generated. The

remnant of the explosion can continue to give off X rays for hundreds or thousands of years. One beautiful example of this type of X-ray object is the Cassiopeia A remnant.

Neutron Stars and Black Holes. Even after a star has blown itself to bits, it may live on in a kind of "life after death." An object known as the Crab Nebula was seen by the ancient Chinese astronomers as a supernova in the year 1054. Yet an X-ray picture taken recently shows a glowing region with a bright point near the center. The point is thought to be the rapidly spinning core of the star that exploded 900 years ago. The part of the star remaining after a supernova like this is called a *neutron star*. It is called a neutron star because the matter left in the star is packed tightly by gravity. It then looks like the nucleus of a giant atom. Not all supernova explosions seem to leave neutron stars at their centers, however. For instance, there is no X-ray point visible in the center of Cassiopeia A. One possible explanation for this is that some explosions do succeed in completely blowing up the star. There is, however, another more interesting possibility.

Suppose that the part of the star left over after an explosion was much bigger than that which makes a neutron star. Then there would be too much gravity for the material to resist shrinking indefinitely un-

Like blinking rays from a lighthouse beacon, this bright neutron star at the center of the Crab Nebula sends out bursts of radiation in beams.

© 1980 Smithsonian Institution



der its own crushing weight. What would happen then would be the formation of an object called a *black hole*.

The black hole is perhaps the most far-fetched object theoretical physicists have yet conceived. Nevertheless, X-ray astronomers think they may have found several in the skies. The first black hole candidate, Cygnus X-1, is an incredibly strong X-ray emitter that does not fit into any of the known categories of X-ray objects. It does, however, show the properties physicists associate with black holes. (See "Black Holes" on page 240.)

Galaxies and Quasars. Galaxies are groups of billions of individual stars. Our own Milky Way galaxy has over 100,000,000,000 stars in it. As we look with optical telescopes beyond our own galaxy, we can see that the universe is filled with billions of galaxies that differ in size, shape, and brightness.

X-ray views of galaxies show a range of properties. Nearby galaxies like our galactic neighbor, the Great Nebula in Andromeda, have dozens of individual X-ray stars similar to the ones we've observed in our galaxy. Other galaxies, too distant for their individual X-ray stars to be seen separately, seem to produce more X rays than would be expected from adding up the contributions of each X-ray star. Still other galaxy-like objects (Seyfert galaxies) generate much more X radiation than can be explained as coming from individual stars. Something that manufactures huge amounts of energy must be happening in the centers of these objects.

Quasars also give off X rays. Discovered through their radio emissions, quasars are the most distant objects known. They mark the edge of the visible universe. Nearly every deep exposure taken with the Einstein telescope showed powerful X ray-emitting quasars. They are everywhere. In fact, some astronomers now suspect that quasars may produce the X radiation other satellites have detected as diffuse background radiation. (See "Quasars" on page 236.)

Clusters of Galaxies. Optical studies of galaxies have shown that there is a larger

scale of structure in the universe. Hundreds or thousands of galaxies are bunched together by gravity into structures called clusters of galaxies. We have seen that some galaxies emit more X rays than the sum of X-ray emissions from their component stars. Some clusters of galaxies are also too bright in X rays to be explained by emissions from individual galaxies. Something else is generating the X rays.

The discovery that clusters give off more X rays than expected is one of the most important discoveries of X-ray astronomy. Many astronomers believe that the X rays originate from vast amounts of gas hiding in the reaches between the galaxies. The gas, undetectable by any other means, emits X rays because it is at extremely high temperatures.

This X ray-emitting gas within clusters may be the "missing mass" astronomers have long sought. The missing mass can provide an answer as to whether the universe is closed or open. If there is sufficient mass in the universe, the universe will one day stop expanding. It will reverse itself and come crashing back together. If, on the other hand, there is not sufficient mass the universe may be open, expanding forever.

WHAT LIES AHEAD

X-ray astronomers are looking forward to the day when they will be able to make observations continuously, as ground-based astronomers do. No X-ray satellite yet launched has lasted for more than a few years. This means that valuable observations grind to a halt each time a satellite quits working.

The advent of NASA's space shuttle will probably make it possible to maintain an observation platform in space indefinitely. The first such space platform will be the Optical Space Telescope. Some years later (before 1990, it is hoped) an Advanced X-Ray Astronomy Facility will be carried into orbit. It will have an enlarged and improved version of the Einstein telescope and will provide sharper pictures of the more distant X-ray objects. Astronomers will then be able to find new answers—but will inevitably find still newer questions.



©1961, by the California Institute of Technology and the Carnegie Institution of Washington

Huge nebulae are among the most interesting of celestial structures. They are believed to serve as the birthplace of stars. The photo above shows the Lagoon Nebula in the constellation Sagittarius.

THE STARS

by Cecilia Payne-Gaposchkin

The thousands of millions of stars in the heavens are so far away that, with the exception of our sun, they are visible only as points of light. The sun alone shows a distinct disk, and we can make out many details of its surface. As a result, astronomers have been able to learn much more about it than about any of the other stars. The sun is a typical star, differing from most others only in scale. Hence a brief survey of its features will help answer the question: "What sort of body is a star?"

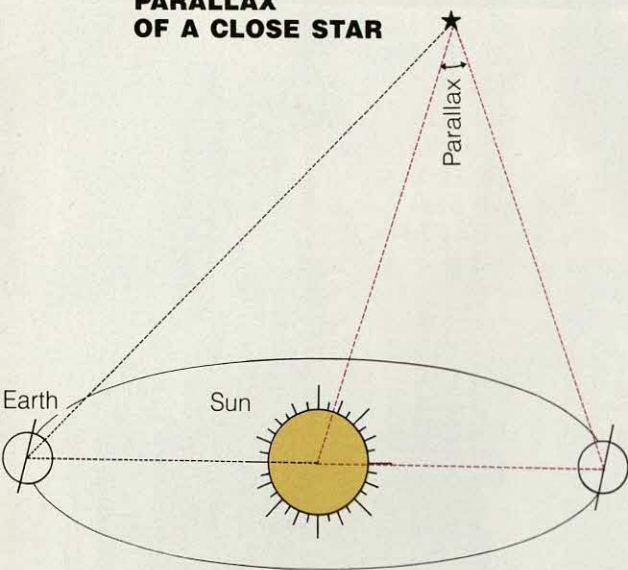
The sun is a globe of glowing gas. Despite its apparently sharply defined disk, it has no solid surface. Its diameter is about 109 times the earth's, and its volume is more than a million times that of our planet.

However, it is much less dense, and its mass is only 330,000 times that of the earth. Estimates of its surface temperature range from 4,100° Celsius to 5,500° Celsius. The temperature of the center may reach 16,000,000° Celsius.

The elements present in the gases of the sun occur also on the earth. The lightest of these elements, hydrogen, is by far the commonest. Roughly speaking, the heavier elements are progressively less abundant.

The surface of the sun shows a delicately mottled and swiftly varying pattern. Large gaseous clouds hang poised over it. Sometimes they shoot off into space. More often they cascade downward like fountains. These are the sun's prominences,

PARALLAX OF A CLOSE STAR



Once the parallax of a star has been determined, the star's distance from the earth can be calculated.

whose motions are probably governed by magnetic forces. Sudden, brilliant flares occur on the surface and send out sprays of atoms into space. Sometimes the solar surface shows sunspots, which look like dark holes. Actually they are very bright. Their apparent darkness is an effect of contrast with the still more brilliant surface nearby. There are strong magnetic fields within the spots. The sun as a whole, however, has a very small magnetic field, if any.

When a total solar eclipse takes place, it is possible to make out the sun's outer envelope, which is generally obscured by the greater brilliance of the disk. Close to the edge is seen a rim of rosy light—the chromosphere, which derives its color from the hydrogen it contains. Farther out stretch the pearly streamers of the corona. The form of these streamers suggests that, like the prominences, they are governed by magnetic fields.

The superficial properties of the sun—its high temperature, its gaseous spectrum, its ever-changing surface, perhaps even its prominences—are probably common to most stars. Since all stars except the sun appear only as points of light, we cannot directly observe the details of their sur-

faces. But astronomers have gained a surprising amount of information about these heavenly bodies by indirect means.

The modern view is that these heavenly bodies are immensely varied and that they are all evolving in accordance with a definite pattern and at differing rates. We can estimate the speed and direction of these changes. We can assign a date to the origin of the stars. We can show that all are not of the same age. We can estimate their span of life. In the following pages we shall sketch the results that have already been obtained.

THE NUMBER OF STARS

About four thousand stars are visible to the unaided eye. Even a small telescope reveals hundred of thousands. Perhaps twenty thousand million could be photographed with a 500-centimeter telescope. Can we suppose that the numbers would continue to mount if more and more powerful telescopes were built? Would we ever come to a stopping place?

The answer is a complex one, for space is not filled uniformly with stars. The universe, as we know it, consists of a number of aggregations, or groups, of stars. The stars that we see with the unaided eye, and most of those that we photograph, are parts of a great system of stars, known as the Milky Way system—our galaxy. It is very large. Light, traveling at a rate of 300,000 kilometers a second, takes perhaps 100,000 years to cross it. Yet it is finite. Our galaxy is a flattened, disk-shaped system, whirling like a pinwheel. It consists of a densely populated center wreathed about with coiled spiral arms, rich in stars, gas, and dust. It contains about a hundred thousand million stars.

Our system is not the only one. We can observe many other galaxies, some very like our own, some different. There are perhaps a thousand million observable galaxies. Our own system contains more stars than the average. Perhaps we shall not be far wrong if we estimate that all the observable galaxies taken together contain about ten million million million stars. Most of these stars are so distant that we cannot

Most stars occur in pairs or clusters. At right, a galactic, or open, cluster—the double cluster in Perseus.



Lick Observatory

hope to see them individually. The collective light of all the stars in a distant galaxy is observable only as a faint blur.

It would be impossible to make a list of all observable stars. The first partial list was drawn up by the Greek astronomer Hipparchus over 2,000 years ago. It has not come down to us. The first extant star catalogue is to be found in the *Almagest* of the Alexandrian astronomer Claudius Ptolemy, who lived in the second century A.D. Johann Bayer, a German astronomer, published a famous star atlas, *Uranometria*, in 1603. Since that time many star catalogues have been compiled for a variety of purposes. One of them describes the spectra of the quarter of a million brightest stars. Others contain lists of double stars, or of stars that vary in brightness, or of stars whose positions, or distances, or motions have been accurately measured.

NAMING STARS

Stars are designated in various ways. The very brightest ones—about fifty in all—have been given specific names. Some of these names, such as Pollux, Capella, and Sirius, are of Greek or Latin origin. Others, including Rigel, Betelgeuse, Algol, and Altair, come from the Arabic.

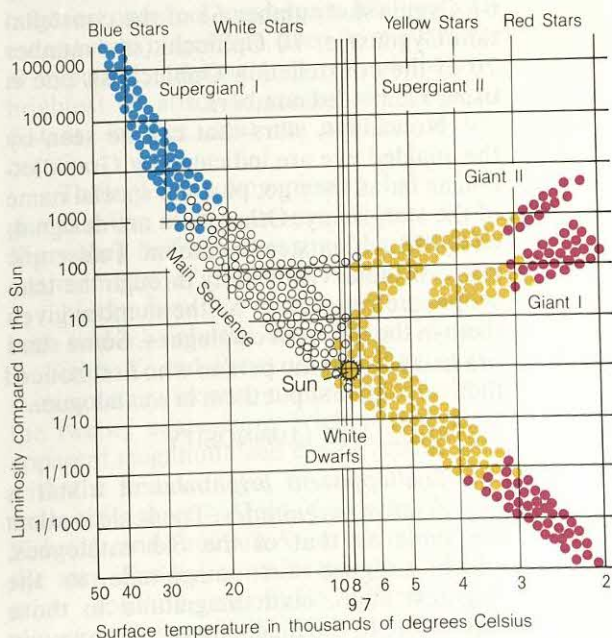
In his *Uranometria*, Johann Bayer introduced the method of using Greek letters for the brightest stars of a constellation.

This method is still employed. The Greek letter is followed by the possessive case of the Latin or Latinized form of the constellation. Thus Alpha (α) Orionis is a bright star in the constellation Orion.

Another method was first used by the

The Hertzsprung-Russell diagram categorizes stars according to magnitude and spectral class (color).

THE HERTZSPRUNG-RUSSELL DIAGRAM





©California Institute of Technology and Carnegie Institute of Washington

Stars are separated by interstellar material and often obscured in nebulosity. Above: the Horsehead Nebula, a dark cloud of interstellar material in Orion. Left: the Great Nebula, seen with the naked eye as the middle "star" in Orion's sword.



English astronomer John Flamsteed in his *Historia Coelestis Britannica* (1725). He numbered the stars of a constellation from west to east across it. When one refers to 61 Cygni (star number 61 of the constellation Cygnus) or 70 Ophiuchi (star number 70 of the constellation Ophiuchus), one is using Flamsteed numbers.

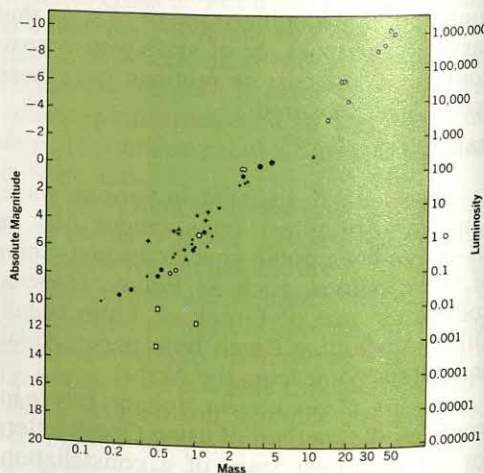
Nowadays, stars that can be seen by the unaided eye are indicated by Greek letters, as far as they go, plus the special name of the star, if any. Other stars are designated by their Flamsteed numbers. Telescopic stars, or those visible only through the telescope, are referred to by the number given them in the different catalogues. Some stars are named after the person who first noticed them or who first put them in a catalogue.

DISTANCE AND LUMINOSITY

The *apparent brightness* of a star is expressed in *magnitudes*. The scale is about the same as that of the old catalogues, which assigned first magnitude to the brightest stars, sixth magnitude to those just visible to the unaided eye. At any point

in the scale, a given difference of magnitude corresponds to a definite ratio in brightness. The scale is such that if one star is five magnitudes brighter than a second star, it is 100 times as bright. This is true, whether the respective magnitudes are 1 and 6, or 16 and 21. A few stars have been found to be a little brighter than first magnitude. These stars are assigned negative magnitudes. For example, Sirius is of magnitude -1.52 . The brightest star in the sky is, of course, the sun, with magnitude -26.72 . The faintest stars observable at present are of about magnitude 23. Thus the observed range is nearly fifty magnitudes, corresponding to a ratio of one to a hundred million million.

Mass-luminosity relation: usually the more massive a star is, the brighter it is. This graph plots the brightness of a number of average stars against a logarithmic scale of their masses with the sun's luminosity and mass being equal to 1.

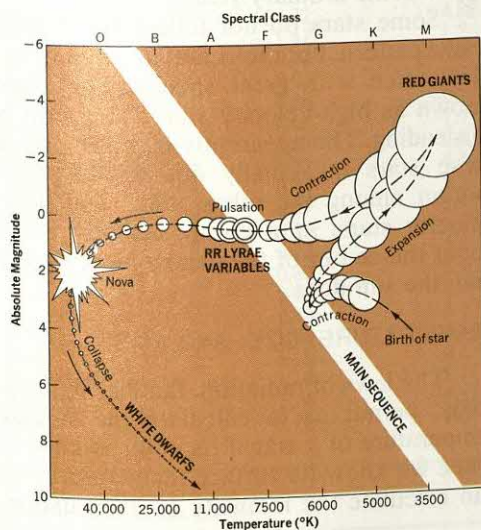


These are the magnitudes of the stars as we see them. A bright star may have a "bright" magnitude because, like the sun, it is close by—actually the sun is not a particularly bright star—or because it is really very bright. We cannot tell which, however, unless we also know how far away a star is.

The distances of certain stars can be measured by the same principles as those used in surveying on the earth. In either case we use triangulation. As the earth moves around the sun in its orbit, it describes a circle (really an ellipse that is very nearly a circle) 300 million kilometers in diameter. Suppose we observe a star on a given date and again, six months later, when the earth is on the opposite side of its orbit. If the star is close enough to the earth, it will appear to be slightly displaced against the background of more distant stars.

We then measure the angle formed by two lines drawn from our two positions on the earth's orbit to the star. Thus we obtain the parallax of the star, which is equivalent to one half the angle that we have just mea-

Typically, when cosmic material condenses to form a star, the star will fall on the H-R diagram's main sequence. Near the end of its life, the star expands as a red giant, then contracts again, sometimes erupting as a nova. Finally it becomes a white dwarf and dies.



sured. Once we know the parallax of a star, we can easily determine its distance from the earth by trigonometry. Only a few thousand stars are close enough to us to be measured directly by the method we have just described. The parallax of other stars or groups of stars must be measured in more indirect ways. We discuss these elsewhere.

When the distance of a star is known, we can calculate from its apparent brightness, or apparent magnitude, what its brightness would be at some standard distance. This is its *absolute magnitude*. The standard distance that has been chosen is ten parsecs. A parsec is the distance at which the parallax of a star is one second of arc (parsec = *parallax* + *second*). Since there are 360 (degrees (°)) in a circle, and each degree is divided into 60 minutes ('), and each minute into 60 seconds ("), there are 1,296,000 arc seconds in a circle ($360 \times 60 \times 60$). Thus one second of arc is $1/1,296,000$ of the circumference of a circle. If the sun were at a distance of ten parsecs, its magnitude would be about 4.86. This represents its absolute magnitude.

Table 1 lists the twenty brightest stars, from the viewpoint of apparent magnitude. The absolute magnitude is also given in each case. Notice that the sun, with the brightest apparent magnitude, has the faintest absolute magnitude. Deneb, the faintest in apparent magnitude, is one of the two brightest in absolute magnitude. It is so distant, in fact, that its parallax is uncertain.

Table 2 is a companion to Table 1. It gives a list of the twenty nearest stars. More than half the stars given in this table are members of double or triple systems.

Notice how few stars are common to both tables—only the sun, Alpha Centauri A, Sirius A, and Procyon A. Many of the stars are very inconspicuous. Only eight of the twenty nearest are brighter than sixth apparent magnitude and can be readily seen with the unaided eye. In absolute magnitude only three—Alpha Centauri A, and Sirius A, and Procyon A—are brighter than the sun. The list contains no stars of negative absolute magnitude. There were eleven such stars in Table 1.

The table of the nearest stars gives a good idea of what the commonest stars are like. The majority of them are fainter in absolute magnitude than the sun, many of them very much fainter. The faintest star known is about $\frac{1}{400,000}$ as bright as the sun. The brightest nonvariable star known is about 400,000 times as bright. Our own star, therefore, is just about average in this respect. However, the number of stars that are fainter than the sun is far greater than the number of those that are brighter.

MOTIONS OF THE STARS

The heavens seem to be unchanging, and we commonly talk of the "fixed stars." All the stars, nevertheless, are in motion. Only their remoteness and our short span of observation give the illusion of fixity. The measured speeds of many stars through space are very large by human standards. Some stars are traveling at hundreds of kilometers a second.

The motions of stars may be measured by precise comparisons of their positions from year to year. However, even the fastest-moving stars do not move as much as a degree in a thousand years. Only the precision of modern instruments enables us to observe the change of position and to predict the shapes of such now-familiar constellations as the Big Dipper in 100,000 years or so. The spectrograph, for example, measures speeds toward us and away from us.

When the information obtained from changes of position and from spectrographic measurement is analyzed, we find that the motions of stars, like their physical dimensions, are not distributed at random. The most obvious apparent regularity is that stars on one side of us appear to be approaching, while those on the opposite side appear to be receding. This is caused by the changing direction of the earth as it goes around the sun.

Stellar motions show some regularities. There is a persistent pattern that was long ago called *star streaming*. We now know it is the consequence of a regular rotation of our whole stellar system, the Milky Way



U.S. Naval Observatory

A time exposure of the starry sky above the North Pole. The apparent circular trails are the result of the earth's rotation on its axis.

system or galaxy, about its center. The stellar traffic is, so to speak, going around a gigantic rotary. The sun is swept along with the rest at about 190 kilometers a second. The sun takes about 200 million years to make the circuit. This interval is sometimes called the *cosmic year*, by analogy with the ordinary year.

Some stars do not follow the regular traffic pattern, but cross it at all sorts of angles, often with great speeds. These are known as high-velocity stars. The term is misleading. The apparently high velocity of these stars is actually due to the sun's motion among the other stars. The high-velocity stars, as a group, actually move about the center of the galaxy more slowly than the sun does.

TEMPERATURE, SIZE, AND MASS

The laws of radiation, familiar to physicists, permit us to calculate the surface temperature of a star if its color is known. Once we know its surface temperature, we can calculate the intensity of its radiation.

The amount of energy radiated per square unit is proportional to the fourth power of the absolute temperature. Absolute temperature is based on the lowest possible temperature, "absolute zero." Absolute zero equals -273.16° Celsius. This law of radiation permits us to find the sizes of stars of known surface temperature and absolute magnitude.

The absolute magnitude tells how much energy the star is emitting per second. The temperature tells how much energy is being emitted per square unit per second. The ratio of these two quantities gives the star's surface area. We can then find its diameter.

The mass of a star can be determined by the gravitational effect of the star on another object. The only stars whose masses can actually be measured in this way are double stars. Almost all stars of known mass conform to the rule that the stars of brightest absolute magnitude are the most massive. This is the mass-luminosity law.

Surface temperatures, absolute magnitudes, and diameters for a number of typical stars are given in Table 3. Masses and densities are also included.

The stars are arranged in four groups. The *supergiants* are very large and of low density. The *giants* are smaller, but still large and diffuse. The *main-sequence stars*, which include our own, are smaller still. The *white dwarfs* are exceedingly small and dense. Within each group the bluest stars are the smallest and densest.

The subdivisions of Table 3 are not arbitrary, for the properties of stars are not distributed at random. Most of the stars in the neighborhood of the sun belong to the main sequence. This is a regular series that runs from hot, bright, rather dense stars, such as Hadar, down through cool, faint, small, dense stars, such as Krueger 60 A. The faint end of the sequence is much more thickly populated with stars than the bright end. For every star like Hadar Beta Centauri, there are millions like Krueger 60 A.

| Table 1 THE TWENTY BRIGHTEST STARS | | | Table 2 THE TWENTY NEAREST STARS | | | |
|---------------------------------------|-----------------------|-----------------------|--------------------------------------|-----------------------|-----------------------|----------------------------|
| STAR | APPARENT MAGNITUDE | ABSOLUTE MAGNITUDE | STAR | ABSOLUTE MAGNITUDE | APPARENT MAGNITUDE | DISTANCE IN LIGHT YEARS |
| Sun | -26.72 | 4.86 | Sun | -26.72 | 4.86 | about 8 minutes |
| Sirius* | -1.47 | 1.2 | Proxima Centauri* | 11.0 | 15.40 | 4.3 |
| Canopus | -0.73 | -4.5 | Alpha Centauri A* | 0.33 | 4.72 | 4.3 |
| Vega | 0.04 | 0.6 | Alpha Centauri B* | 1.70 | 6.09 | 4.3 |
| Arcturus | 0.06 | -0.1 | Barnard's Star* | 9.5 | 13.18 | 6.0 |
| Rigel | 0.08 | -7.0 | Wolf 359 | 13.5 | 16.64 | 7.7 |
| Capella* | 0.09 | 0.3, 0.6 | +36° 2147 | 7.5 | 10.47 | 8.3 |
| Alpha Centauri A | 0.33 | 4.7 | Sirius A* | -1.5 | 1.26 | 8.7 |
| Procyon* | 0.34 | 2.1 | Sirius B* | 8.4 | 11.26 | 8.7 |
| Achernar | 0.47 | -1.6 | Yale 343.1* | 12.5 | 15.30 | 9.0 |
| Hadar | 0.59 | -5.1 | — | 13.0 | 15.80 | 9.0 |
| Altair | 0.77 | 2.4 | Ross 154 | 10.6 | 13.26 | 9.6 |
| Betelgeuse | 0.80 | -5.8 | Yale 5736 | 12.2 | 14.70 | 10.3 |
| Aldebaran | 0.86 | -0.3 | Epsilon Eridani | 3.8 | 6.21 | 10.8 |
| Spica* | 0.96 | -3.5 | Yale 5475 | 12.2 | 14.59 | 10.9 |
| Antares | 1.08 | -4.5 | Yale 2730 | 11.0 | 11.72 | 10.9 |
| Pollux | 1.15 | 0.2 | 61 Cygni A* | 5.2 | 7.54 | 11.1 |
| Fomalhaut | 1.16 | 1.7 | 61 Cygni B* | 6.0 | 8.34 | 11.1 |
| Beta Crucis | 1.24 | -4.25 | Epsilon Indi | 4.7 | 6.98 | 11.4 |
| Deneb | 1.26 | -7.0 | Procyon A* | 0.5 | 2.76 | 11.5 |
| *member of multiple system | | | *member of binary or multiple system | | | |

Table 3

TEMPERATURES AND DIMENSIONS OF TYPICAL STARS

| STAR | | SURFACE TEMPERATURE IN DEGREES C | RADIUS IN SUNS | MASS IN SUNS | DENSITY IN SUNS | ABSOLUTE MAGNITUDE |
|---------------------------|-------------------|--|----------------------|-----------------|-----------------------|-----------------------|
| SUPER- GIANTS | Beta Lyrae | 12,000 | 19.2 | 9.7 | .0014 | 0.1? |
| | Rigel | 11,300 | 78 | 20? | .00004 | - 7.0 |
| | Deneb | 10,200 | 96 | 20? | .00002 | - 7.0 |
| | Gamma Cygni | 4,100 | 67 | 20? | .00007 | - 5.0 |
| | Betelgeuse | 2,700 | 1000 | 10? | .0000005 | - 5.8 |
| | Antares | 2,900 | 776 | 20? | .00000004 | - 4.0 |
| GIANTS | Capella | 4,800 | 13 | 2.1 | .00096 | 0.3 |
| | Arcturus | 3,800 | 35 | 8 | .00018 | - 0.1? |
| | Aldebaran | 2,700 | 87 | 4 | .000006 | - 0.3 |
| | Beta Pegasi | 2,000 | 40 | 9 | .00014 | - 1.4 |
| MAIN SEQUENCE STARS | Hadar | 21,000 | 22 | 25 | .0023 | - 5.1 |
| | MU-1 Scorpii | 20,000 | 5.2 | 14.0 | .1000 | - 5.1 |
| | Sirius A | 10,200 | 1.9 | 2.3 | .335 | 1.2 |
| | Altair | 7,300 | 1.6 | 1.7 | .415 | 2.4 |
| | Procyon A | 6,800 | 2.6 | 1.8 | .102 | 2.1 |
| | Sun | 5,900 | 1.0 | 1.0 | 1.0 | 4.86 |
| | 61 Cygni A | 2,600 | 0.7 | 0.58 | 1.69 | 7.65 |
| | Krueger 60 | 2,800 | 0.35 | 0.27 | 6.30 | 11.9 |
| | Barnard's Star | 2,700 | 0.15 | 0.18? | 53.3 | 13.2 |
| | | | | | | |
| WHITE DWARFS | Sirius B | 5,000? | 0.022 | 0.99 | 90,000 | 11.4 |
| | 40 Eridani B | 5,000? | 0.018 | 0.41 | 71,000 | 11.2 |
| | Van Maanen's Star | 5,000? | 0.007 | 0.14? | 47,000 | 14.2 |

The giant stars are a different group, all of nearly the same brightness, much larger and less dense than the main-sequence stars. They are as rare as the bright main-sequence stars. The supergiants are stellar freaks, extremely rare in space. They force themselves on our attention because they are so bright, but there are thousands of giants for every supergiant.

The white dwarfs, on the other hand, are so very faint that they are quite hard to find. As a matter of fact, white dwarfs are probably more common than main-sequence stars like our sun.

STAR STRUCTURE

Observation shows that the stars are gaseous on the surface. The laws of physics can be used to probe their interiors, and to show that they are gaseous through to their centers also. Because of the weight of the outer parts of a star, such as the sun, the center is extremely dense. The pressure is

about one million metric tons per square centimeter. It seems probable that all main-sequence stars are similar to the sun in structure.

The interiors of giant stars probably have about the same central temperature as main-sequence stars, but they are greatly distended and their structure is less simple. Supergiants are still more puzzling. We do not understand what sustains their enormous, low-density bulk. The structure of the white dwarfs is radically different. The internal material of these extremely small stars is packed to almost incredible density, and their atoms must be tightly jammed together.

SOURCES OF STELLAR ENERGY

The sun radiates energy at the rate of 3.79×10^{33} ergs a second. An erg is a unit of energy. The source of this energy was long a mystery. Ordinary processes of combustion or contraction could not account for it.

According to modern theory the energy that streams out of the sun is due to reactions involving the nuclei of hydrogen atoms. The hydrogen nuclei are built up into helium nuclei, releasing energy as a by-product. There are probably two ways in which this is done. The hydrogen nuclei may combine directly, or in a roundabout process in which carbon atoms also take part.

The discovery of the source of solar energy has solved the problem for the other main-sequence stars as well, since their internal conditions are very similar to those in the sun. In all these stars, hydrogen is being consumed; helium is the final product; and radiated light a by-product.

EVOLUTION OF STARS

The main-sequence stars probably begin their existence as large, diffuse, cool bodies. Gravitational forces within them cause them to contract. As they contract, they grow denser and hotter inside. When the interior of a contracting star reaches the temperature at which the hydrogen-to-helium reaction can proceed, its contraction is stopped by the outpouring radiation. The star then takes its place on the **main sequence**. If it is of high mass, it will have become rather hot at the surface by this time. A star of lower mass is cooler when it reaches the main sequence. The place where a star enters the main sequence depends on its original mass.

If all main-sequence stars were radiating energy, and therefore consuming hydrogen, in exact proportion to their masses, the process would go on for exactly the same interval of time for all of them. But this is not the case. The mass-luminosity law states that a star's luminosity is proportional, not to its mass, but nearly to the cube of its mass. Thus a main-sequence star with ten times the sun's mass is radiating and using up its hydrogen about a thousand times as fast as the sun. Therefore, the hydrogen in the more massive star will last only one thousandth as long as the sun's.

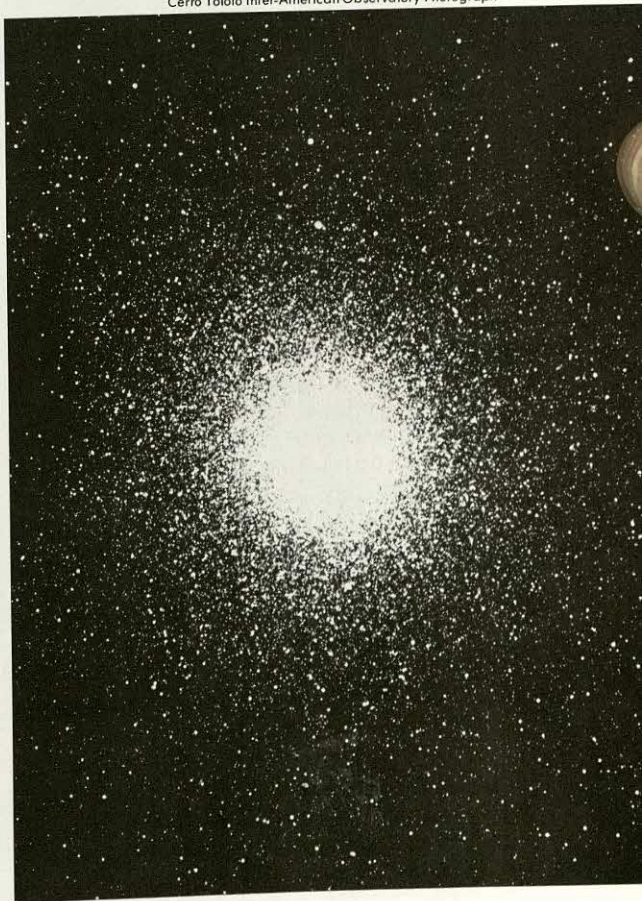
We calculate that, at the sun's rate of consumption, its original store of hydrogen

would last altogether about 10,000,000,000 years. About half this interval is already past. A star with a hundred times the sun's mass would last on the main sequence only about 10,000 years. We do not know of any such massive stars. Perhaps the reason is that they are too short-lived to be observed.

When a star has used up its available supply of hydrogen, energy can no longer be supplied by hydrogen-to-helium reactions. However, the star can release gravitational energy by contracting. As it does so, the temperature of the interior rises again. As a result the star begins to consume its helium atoms. It increases in size and becomes a *red giant*.

47 Tucanae in the constellation Tucana is the brightest globular star cluster known. It is seen only from Southern Hemisphere locations. This photo was taken with a 4-meter telescope at Cerro Tololo Observatory in Chile.

Cerro Tololo Inter-American Observatory Photograph



In time, the nuclear energy resources of the star are used up. What happens next, according to modern theories, depends upon the mass of the star. Stars ranging from very small to about 1.4 times the mass of the sun will gradually contract and grow fainter and denser. For brief periods they may become novae, which are discussed later. Eventually they wind up as *white dwarfs*. They will continue to glow until they can contract no longer. Then they will cease to give off light.

More massive stars, however, will undergo much greater gravitational collapse. They will become *neutron stars*—stars whose subatomic particles have condensed into neutrons, tightly packed together. Such stars may be only a few kilometers wide. Very massive stars may collapse completely and become *black holes*—stars that are so dense that light cannot escape from their surface.

STAR GROUPS

Very few stars are completely isolated. Almost all occur in groups. Pairs of stars are common. Among the twenty nearest stars, there are many pairs. There is also a triple—that is, a three-star group. Stars also tend to form clusters. Certain clusters, such as the Pleiades, contain a few hundred stars. Others, like the Great Cluster in Hercules, have thousands or tens of thousands of members. The important thing to realize about groups of stars, whether doubles or clusters, is that they cannot possibly have come together by chance. They must have been formed together.

The members of a star cluster started out together. Their condition today shows how each one has evolved. The members of a young cluster of stars might be lined up along the main sequence. At first sight, the Pleiades seems to be such a cluster, for it shows a progression of stars, from the brightest, which are also the bluest, down

to the faintest, which are smaller, redder, and denser. Actually, the very brightest Pleiades have already moved a little way off the main sequence. The Hyades, visible to the eye as a V-shaped group of stars in the constellation Taurus, are farther advanced. Several of their members have actually become red giant stars. The longer a cluster has been developing, the less of the primitive main sequence is left. There are also clusters whose faintest members have apparently not reached the main sequence yet. They are still contracting.

The older a star cluster, the less of its main sequence is left. Some very young clusters have only partially formed their main sequences. The ages of the clusters like the Pleiades range from about a million years for the youngest to nearly a thousand million for the oldest.

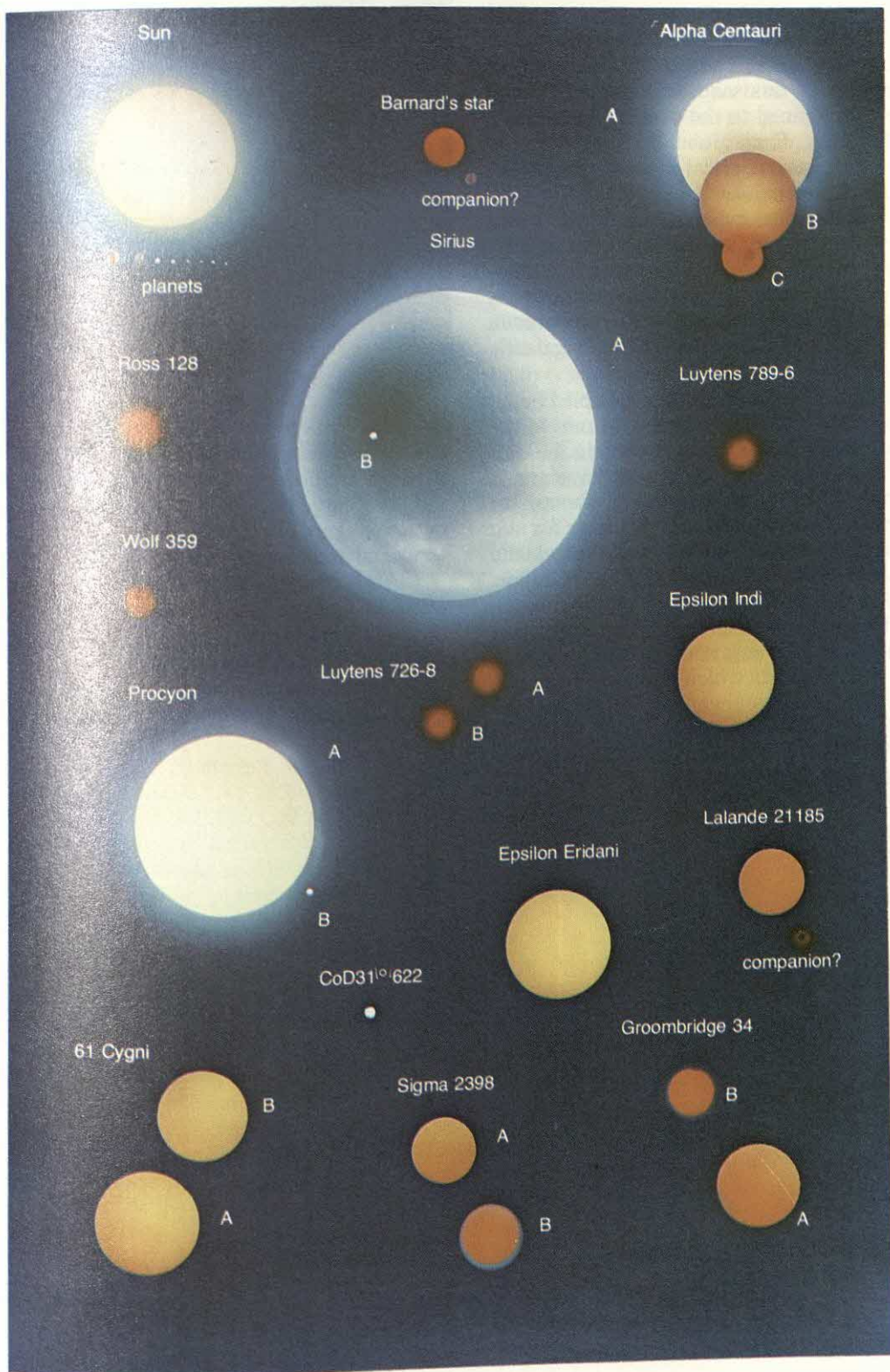
THE MILKY WAY

When we spoke of the number of the stars, we described the disk-shaped pinwheel, isolated in space, that we call our Milky Way system. Vast numbers of stars are crowded toward the center. There are also spiral arms made up of stars, gas, and dust. These arms are not easy to make out within our own system, for we are too near to them. They are easily seen in other distant galaxies. With the aid of the radio telescope we are now determining the position they occupy in our galaxy.

The Pleiades-like clusters all lie in the spiral arms. They tend to be enmeshed in gas clouds and in dust. They are called galactic clusters. They lie in the central plane of the Milky Way, where the stars are thickest. The galactic clusters are typical members of the spiral arms. Many other stars that inhabit the spiral arms are similar in character and probably also in history and age. The spiral arms are continually reborn. They are being formed from interstellar dust and gas.

ASTRONOMY Magazine painting by Mark Paternostro

Scientists estimate that half of all stars have planetary or stellar companions. The farther away a star is, the harder it is to identify its companions. Pictured at right are the relative sizes and spectral differences of our sun and its closest neighbors.



Globular clusters are larger than the galactic clusters. They have thousands or hundreds of thousands of members. They are not confined to the central plane of the Milky Way. Each globular cluster forms a more or less spherical system, which accounts for its name.

Their development patterns show that they are older than galactic clusters. Their ages range perhaps from three to five thousand million years. The colors and sizes of their stars indicate that they are poorer in metallic atoms than the stars in galactic clusters.

We can now trace the probable course of development of the stars in our Milky Way. The globular clusters contain the oldest stars. Obviously they must have arisen first. Their large population shows the immensity of the material available for star formation in the early days of the galactic system. They are poor in metallic atoms and rich in light elements such as hydrogen and helium. We can infer, therefore, that these light elements must have abounded in the material from which the stars formed.

The much younger galactic clusters are much richer in metallic elements. How does this come about? It will be recalled that the nuclear transformations in the stars build light elements into heavier ones. In their final stages, they must be producing metallic atoms, such as iron. The most massive stars run through these stages rapidly. It seems likely that in their final decline, they become unstable and unload their excess mass by exploding. The material that is thrown into interstellar space by such explosions is rich in metallic atoms. It is this material that furnishes the stuff from which fresh stars and clusters of stars are formed. We can show that the loose gases and dust have a tendency to collect in the central plane of the Milky Way. Therefore young stars will tend to be born in this plane. As we have seen, that is just where the galactic clusters are formed.

EXPLODING VARIABLE STARS

Certain stars show more or less notable variations in brightness and sometimes in other respects. They are known as *vari-*

able stars. There are thousands of them in our own galaxy—the Milky Way—and in other island universes, as galaxies are sometimes called.

Among the most interesting are the exploding stars. They are fairly common in our galaxy. About fifty of them explode every year. They are known as new stars, or novae—an inaccurate term, for only old stars are explosive. Although the stars that become novae are about as bright as the sun, they are much smaller, denser, and bluer, and have advanced farther along the evolutionary path. Stars such as the sun have many thousands of millions of years to go before they reach the nova stage. Long before the sun reaches this stage, if modern theories are correct, it will have become a red giant large enough to engulf our planet entirely. This will happen perhaps in a thousand million years.

An ordinary nova becomes about 100,000 times as bright as the sun during the explosion. However, only the surface area is affected. The outer "skin" of the star is blown off. After a few years the disturbance dies away and the star is found to be little altered.

Some stars explode much more violently. Such stars, called supernovae, become perhaps 100,000,000 times as bright as the sun. Supernovae are, perhaps, massive stars that almost blow themselves to pieces, leaving behind a dense core such as a white dwarf or a neutron star. These explosions occur in our galaxy perhaps once in two hundred years. The Crab Nebula, for example, is the expanding remnant of a supernova that was observed in 1054 A.D. At the center of the Crab Nebula is a tiny star, probably a neutron star, that is all that remains of the original stellar mass. It is also an example of certain objects that have been discovered by radio astronomers—pulsars, which emit strong radio signals at rapid intervals.

PERIODIC VARIABLES

Not all variable stars explode. The commonest variable stars are those that periodically grow brighter and darker. Many of them pulsate or oscillate rhythmically.



Kitt Peak National Observatory

Betelgeuse, or Alpha Orion, is the first star, except the sun, of which scientists were able to obtain an extended image. Above is an enhanced photo showing starspots on the face of the star.

cally. Some stars oscillate once every year or two. The Cepheid variables—stars that follow the oscillation pattern of the bright star Delta Cephei—have a pulsating period of a few days. Other pulsating stars have periods between a day and a couple of hours. The slowly pulsating stars are large and red, the faster ones smaller and blue. The period is related to the density of the star: the square of the period is inversely proportional to the density. The fastest pulsation known is that of a dense blue star that once became a nova. Its period is a little over a minute.

The variations of novae and Cepheid variables are not confined to their brightness. Conditions at their surfaces are changing along with the variations of brightness. A study of their spectra shows that Cepheid variables, for example, change in temperature rhythmically as they vary. They are hottest when they are

Some stars, called supernovae, explode violently. Below: photos of the same section of the sky, taken 13 years apart. The photo on the right shows the emergence of a supernova just below the galaxy at center.

Hale Observatories



brightest, coolest when they are faintest.

The novae and the pulsating stars are not only of interest in themselves. They furnish measuring sticks for the distant stellar systems. It has been found that all novae and supernovae that behave in the same way have the same true or absolute brightness. The absolute brightness of pulsating stars depends on their periods. Those with the longest periods are the brightest. Using such stars, astronomers have determined the distance of a great number of galaxies in which these stars occur.

ERRATIC VARIABLES

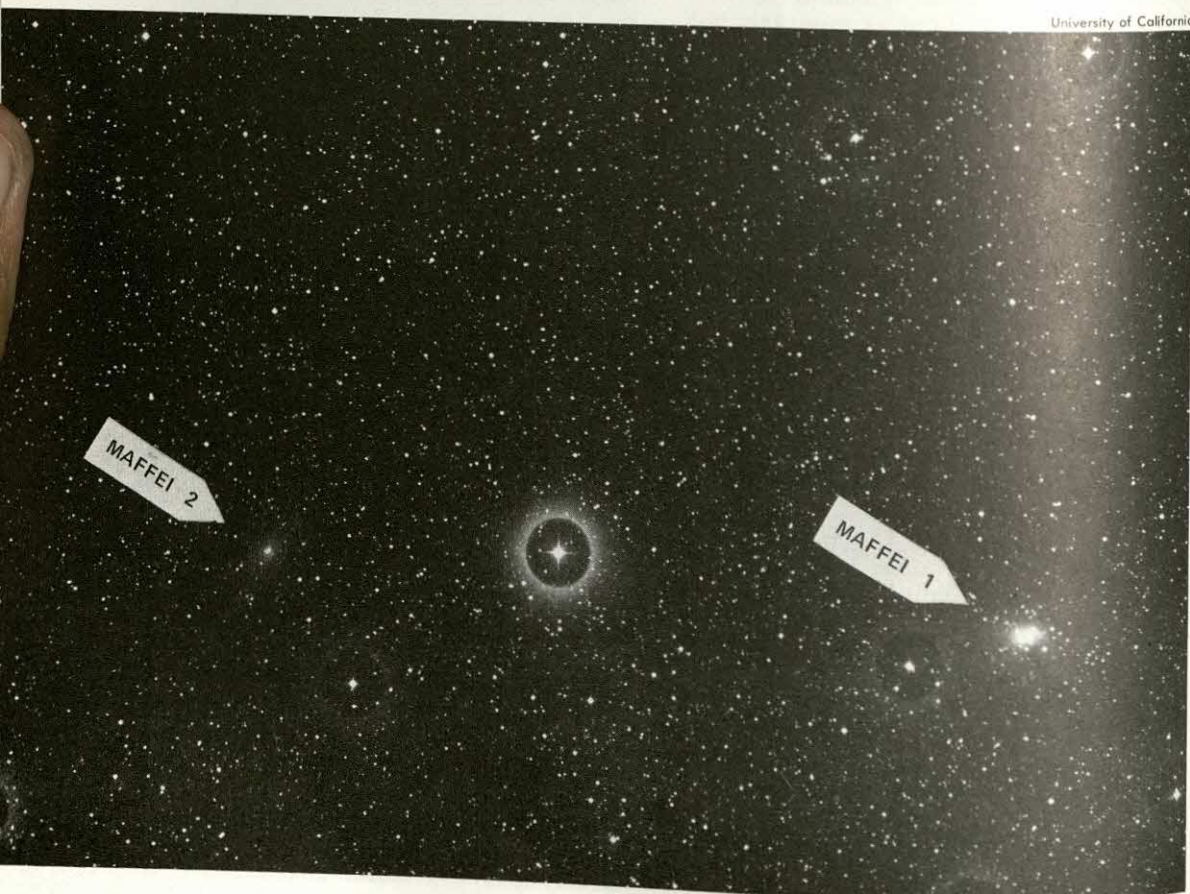
Not all variable stars are explosive or regularly periodic. There is a most interesting group of erratically variable stars, al-

ways found embedded in clouds of gas and dust. They are probably veritable stars in the making, actually contracting out of interstellar material. Some of the extremely young galactic clusters contain many of these stars. Some astronomers think that a supernova explosion led to the formation of such clusters.

Other stars, which vary in brightness little if at all, show remarkable surface variations. A small and fascinating group are the magnetic variables. They display enormous magnetic fields that change polarity with a regular rhythm. The origin and variations of these magnetic fields are somewhat of a mystery. Comparable magnetic fields, which are also not fully understood, are observed within sunspots.

The photo below reveals the cores of two galaxies—Maffei 1 and Maffei 2—amid a veil of interstellar material. It is impossible—even with powerful telescopes—to see the individual stars of these galaxies.

University of California



VARIABLE STARS

Even to the unaided eye, many of the stars display beautiful tinges of color. The intensity of these colors is increased by the telescope, which also reveals the color of many stars that without its aid appear white. Antares is a wonderful ruby red. Somewhat less deep and rich in tone is another red star, Betelgeuse. Aldebaran and Arcturus are red-orange. The sun is, of course, yellow. Procyon is yellowish white; Sirius is white with almost a bluish cast; Mirzam is definitely bluish. These illustrate the color range of all the stars in the sky, red to blue-white, but there are a hundred delicate tints in between—lilacs, amethysts, and greens, forming a beautifully colored pattern in the night skies.

Homer, Cicero, and some other ancient writers described Sirius, the Dog Star, as red, though it is today very clearly white. Can stars, then, change color? They can and do. It has been determined that color is an index of temperature. Red stars are comparatively cool, around 2,000° to 2,600° Celsius. Blue-white stars are the hottest, up to 50,000° Celsius. These are surface temperatures. Inside the stars, temperatures go up to tens of millions of degrees.

Astronomers at one time believed that all blue-white stars were “young” stars and that red ones were “old,” with yellow and orange ones in between. Nowadays the accepted theory is that stars progress from red toward white.

This is what probably happens: a “young” star is a diffuse mass of elements in gaseous form. Its color is reddish, which means it is comparatively cool. But it is, nevertheless, hot enough so that many atoms are excited. They lose electrons. This means, of course, that they lose mass. The star contracts. Contraction creates pressure and greater heat. Nuclear reactions take place, in which some mass is lost but great energy is created. The star grows progressively hotter. Its color changes from red to orange to yellow to white to blue. In its hottest blue-white state a re-

versal sets in. The star starts to cool off, changing color again, growing white, yellow, orange, red and, finally, losing its luminosity, becoming a cold, dead mass. According to this theory of a star's life history, Sirius might possibly have been red during Homer's time though it is white now and at some future date will be red again.

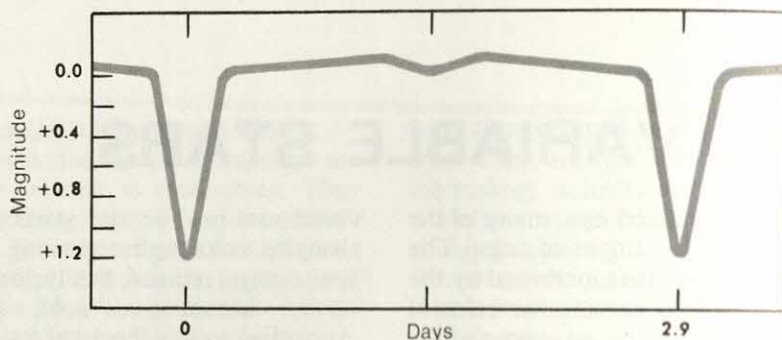
ECLIPSING BINARIES

Sirius also changes periodically in brightness, or apparent magnitude. Thousands of stars do this. They are called *variables*. The variables have been classified according to the reason for their changing brightness. Sirius belongs to the class of *eclipsing variables*. That means that Sirius has a companion star, and the two revolve around a common central point. At times the smaller companion, or Sirius B, gets between Sirius A and the earth, cancelling out some of the luminosity of Sirius A.

Two-star companionships of this sort are called *binaries*. Some of them are very beautiful and impressive when viewed through a telescope. Stars that to unaided vision seem but a single point of more or less doubtful color reveal themselves in the telescope as pairs. These pairs often display amazingly lovely harmonies of color, either contrasted or graded in pleasing nuances.

Algol, perhaps the best-known of the eclipsing binaries, was the first to be discovered by astronomers. Perhaps the ancients suspected that it was out of the ordinary. The name “Algol,” which means “the demon” in Arabic, may have referred to the bizarre quality of the object's light, alternately becoming brighter and then dimming.

In 1667, Montanari discovered that Algol was a variable star, but it was not until 1783 that John Goodricke suggested how its variations occurred. He advanced the theory that what appeared to be a single star was really made up of two: a bright star and a fainter companion that eclipsed the bright star as it revolved



Variations in the brightness of a star can be expressed by a light curve in which the brightness of the star is plotted against time. At right, the light curve for eclipsing binary Algol. This star ranges from optimum brightness of below 0 magnitude to minimal brightness of about 1.2 magnitude over a period of 2.9 days. It is dimmed when its companion star eclipses it.

around it. He noted that the light of Algol becomes dimmer at intervals of about two days and twenty-two hours.

Much has been learned since his day about the two stars that make up the eclipsing binary Algol. The brighter of the two—the primary star—has a diameter three times that of the sun. Its companion is somewhat larger, but is fainter by three magnitudes. The centers of the two stars are about 20,000,000 kilometers apart. Their orbits are almost edgewise to the earth. They are inclined from the edgewise position by only 8° or so.

Another interesting eclipsing binary is Zeta Aurigae. It is made up of a blue star, whose diameter is about seven times that of the sun, and a red star, with a diameter about fourteen times that of its companion. The two stars have approximately the same apparent magnitude. They revolve about each other once every 972 days. They eclipse each other at intervals as they move in their orbits. First we see a partial eclipse, lasting thirty-two hours, then a total eclipse of thirty-seven days, and finally another thirty-two-hour partial eclipse.

The components of certain eclipsing binaries are so close together that they appear as a single star even when viewed through the most powerful telescope. They have been discovered and studied, however, with the spectroscope. Therefore they are called **spectroscopic binaries**.

The two stars of a typical binary of this kind approach the earth and then recede from it in their orbits. The motions

of the stars can be analyzed with the help of the spectroscope because, in accordance with a certain principle in science known as the Doppler effect, the lines in their spectra shift to the violet as they approach the earth and to the red as they recede from it.

The variable stars that we have considered up to this point are called *extrinsic variables*. The changes of brightness that we note in them are brought about by external factors, such as eclipses. We shall now take up the *intrinsic variables*. These are the stars that vary in brightness because of physical changes that take place in them. Among these physical changes are changes in heat output, color, stellar spectrum, and velocity. Certain variables are both extrinsic and intrinsic. They are eclipsing binaries that also undergo internal changes.

NOVAE

A fascinating class of intrinsic variables are the exploding stars known as novae, or "new stars." They were given that name because they appeared to be newly created. We realize now, however, that a nova is not "new." It is a pre-existing star upon which an explosion has taken place.

A faint star, perhaps too faint to be seen by any except the most powerful telescopes, begins to be more brilliant. Within the space of a few days it becomes thousands of times brighter than it was before. The mechanism producing the increase in brightness is most likely the release of an enormous amount of atomic energy—a nuclear explosion. The intensity of the explo-

sion can be measured by the amount of material thrown outwards: the greater the magnitude of the explosion, the greater the amount of material ejected. After some days of maximum brilliance, the luminosity begins to fade. Ultimately, the star is about as faint as it was before. This would seem to indicate that the outburst is far more superficial than it appears to be. The explosion of a nova occurs without warning. On the average, there is an increase of about 60,000 times in brightness, corresponding to 13 magnitudes. The growth of brilliance of a nova is always extremely rapid, and its fading is gradual.

A nova is designated by the word "Nova," followed by the Latin possessive form of the name of the constellation in which it occurs and the year in which the outburst took place. For example, when we refer to Nova Persei 1901, we have in mind the nova that appeared in the constellation Perseus in the year 1901.

The brightest nova of which we have definite record appeared in the constellation Cassiopeia in November 1572. It was observed by the great Danish astronomer Tycho Brahe. Nova Cassiopeiae 1572, to give it its technical name, became so bright that it could be seen in daylight. It faded very gradually and finally disappeared from view in the spring of 1574. There were no telescopes in those days. Otherwise, it would have been visible some time longer. Another particularly brilliant nova was observed by the German astronomer Johannes Kepler in 1604, in the constellation Ophiuchus. It remained visible to the naked eye for about eighteen months. This nova became known as "Kepler's star."

A gaseous envelope develops around a nova. In certain cases, this envelope is spherical. Sometimes it is elliptical; in such instances, it is probably composed of several "shells," each corresponding to a different explosion. Occasionally an envelope attains fantastic size. The one that formed around Nova Aquilae 1918 was something like 1,600,000,000,000 kilometers in diameter after about twenty years. The Crab Nebula in the constellation Taurus shows an envelope that has been expanding for

hundreds of years. It seems to have been formed by a nova that appeared in the year 1054, according to Chinese and Japanese records. Actually, the outburst must have taken place four thousand years or so before 1054, since it would take that length of time for its light to reach us.

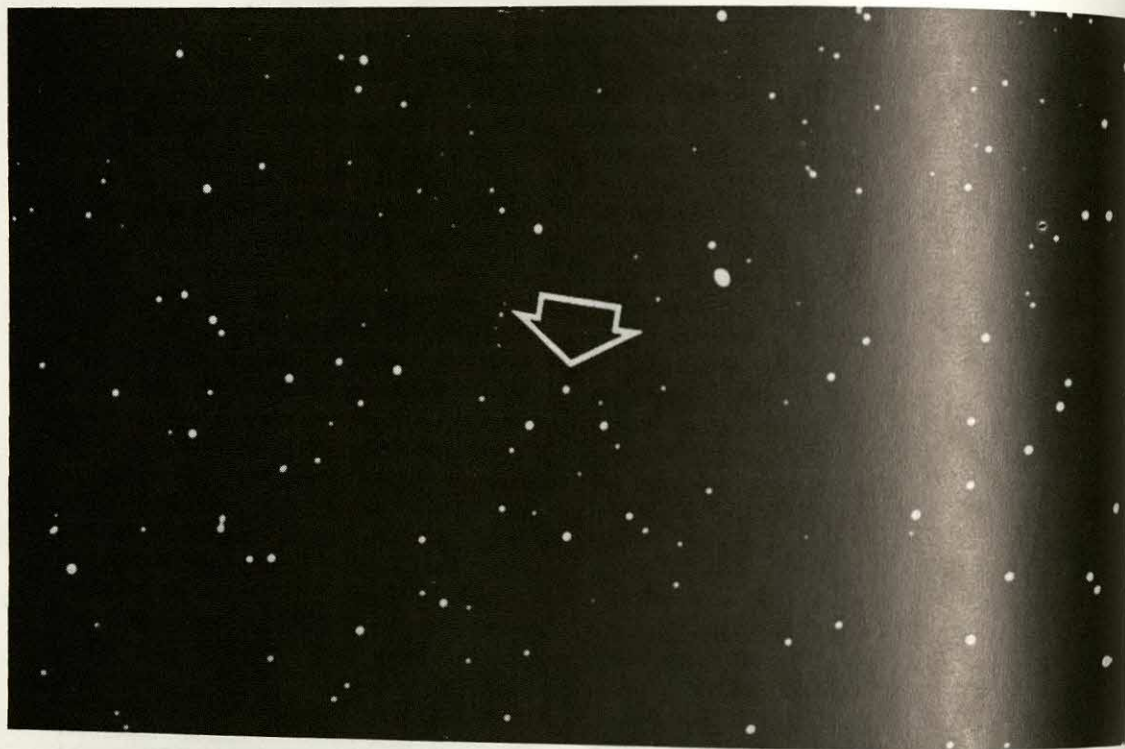
A number of novae explode only once. In others, which are called recurrent novae, two or more outbursts take place. These novae rise to maximum brilliance more slowly than those that explode only once, and they take less time to return to their former faintness. They also have smaller ranges of magnitude than most novae. Among the recurrent novae are Y Coronae Borealis, RS Ophiuchi, and U Scorpii.

We can only conjecture about the cause or causes of the explosions that result in the formation of novae. It was formerly held that they were due to collisions between stars, or between a star and a nebula. Few take this theory seriously now. For one thing, how could we explain recurrent novae on the basis of collisions? We would have to assume that a single star collides again and again with some celestial object in the course of a century. This would be beyond the range of probability.

It is more likely that a nova arises as the result of the sudden release of internal energy. Perhaps this energy release is brought about by a gradual increase in temperature, which finally triggers a nuclear explosion.

SUPERNOVAE

In some cases, the explosions accompanying the rise of novae are particularly spectacular. Such out-of-the-ordinary novae are known as supernovae. The first one recorded by astronomers appeared in 1885 in the central region of the galaxy called the great spiral in Andromeda. It was of the seventh magnitude. It was one-tenth as bright as the spiral itself and ten times as bright as the average nova. At first, astronomers did not consider it as differing appreciably from the other novae known at this time, since the great spiral in Andromeda was thought to be a nebula and much nearer to us than it actually is. Once it was realized



Yerkes Observatory

Some intrinsic variable stars are novae. In the photo above, the arrow points to the star now known as Nova Aquilae 1918 as it appeared before the outburst that made it a nova. On the next page, we see how much brighter the star became after the explosion.

how far away from the earth Andromeda is, it became obvious that this particular nova was an extraordinary one.

We pointed out that the explosion that causes the average nova to become so brilliant does not have relatively far-reaching effects. It is estimated that a nova throws off only one thousandth, or even less, of the matter that it contains. The supernova is affected much more by an explosion. As far as we can judge, it throws off the greater portion of its mass. Furthermore, it does not return to its former state after the explosion takes place. If we adopt the internal-energy theory to account for the rise of supernovae, the energy release must be far greater than in the case of novae.

PULSATING VARIABLES

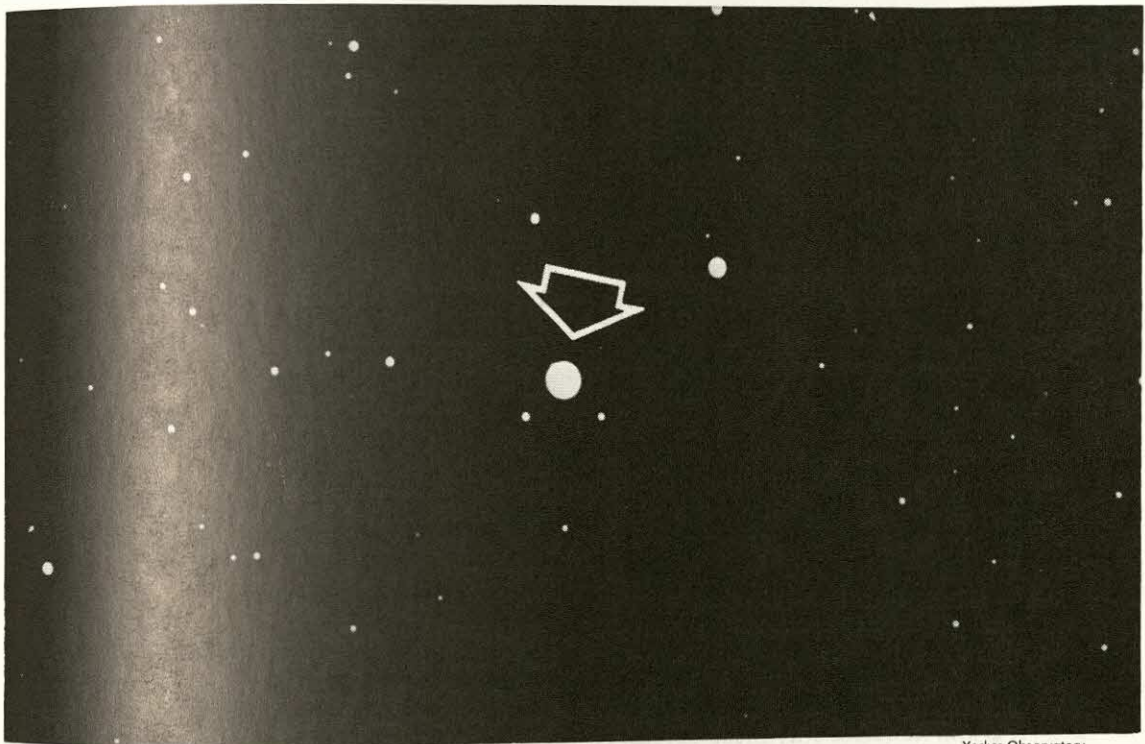
The pulsating variables make up another important group of intrinsic variable

stars. They have received the name "pulsating" because they grow alternately brighter and fainter according to a more or less definite pulse, or rhythm. The pulsating variables may be divided, rather arbitrarily, into two groups: long-period and short-period. In the first group, the period from one peak of brightness to the next ranges from a hundred days to a thousand. In the second group, it ranges from less than one day to about fifty.

LONG-PERIOD VARIABLES

The stars with the longest period are the large red giants, with low density and low surface temperature. Variations in brightness are not altogether regular. These stars are all fairly bright when they reach their maximum brilliance and have conspicuous variations. Hence they can be studied by the amateur astronomer with a small telescope or even a pair of binoculars.

The long-period star called Mira, in the constellation Cetus, is the prototype of the long-period variable. As a matter of fact, the name "Mira stars" is sometimes given to them. Mira was the first long-period vari-



Yerkes Observatory

able to be discovered. It was first observed by David Fabricius, a German theologian and astronomer, in 1596. He thought he had found a nova. It was seen again in the year 1638. The astronomers of that time also assumed that it was a nova, at first, but they soon became aware of its true character as a periodic variable.

Mira has a period of about 332 days. Neither the maximum nor the minimum brilliance in a given cycle can be determined beforehand with any precision. Either may be hastened or delayed by a week or two and occasionally by as much as forty days. There are also considerable variations in intensity of light. The maximum brightness of the star has been known to reach almost the first magnitude. In 1779, Sir William Herschel noted that Mira was almost equal in brilliance to Aldebaran, a first-magnitude star. At other times, however, Mira fails to attain the fifth magnitude at the peak of its brightness.

Its minimum brilliance is more uniform, on the whole, but it, too, is subject to irregularity. It usually falls between the eighth and tenth magnitudes, but it has

been known to be considerably fainter than this limit. In 1783, for example, Sir William Herschel could find no trace of the star in a telescope in which all stars down to the tenth magnitude were visible.

If we consider the whole range within which Mira has been known to vary, we find that the light it emits at certain periods is about ten thousand times as great as that emitted at other times. In its usual range of brilliance, however, it is from twelve to fifteen hundred times brighter at maximum than at minimum.

In spite of its variations, we can mark a certain pattern—an average of waning and waxing. At intervals of about eleven months, the star begins to brighten. The progression from minimum to maximum brightness takes something like a hundred days. Mira retains its brilliance for some weeks. Then it subsides to its former level of brightness. It takes the star almost twice as long to pass from maximum to minimum brightness as from minimum to maximum. Considering the average of its performance, it remains at a high level of brilliance for about two months and at a low one for

about four. It spends the rest of the time changing from one level to the other.

Another famous long-period variable is Chi Cygni. This striking star, which is of a beautiful scarlet color, is more than six thousand times as bright when it is at maximum brilliance than when it is faintest. It waxes and wanes in periods of about four hundred days through a range of eight magnitudes.

Mira and Chi Cygni show exceptionally wide ranges of fluctuations. The average amount of variation in seventy-five long-period variables that were kept under close observation at Harvard University was found to be five magnitudes. This means that the average long-period variable gives out a hundred times more light at maximum brightness than at minimum.

Spectroscopic examination has shown that the increase in the brilliance of long-period stars is caused by periodic outbursts of incandescent gases, chiefly hydrogen. It is still not clear why these outbursts take place.

SHORT-PERIOD VARIABLES

The intrinsic variables with the shortest periods are called the RR Lyrae stars, because the first of the group to be studied was RR Lyrae, the star named RR in the constellation Lyra. They have been discovered in considerable numbers in the globular clusters. For this reason they are sometimes known as cluster-type variables. Their periods range from about an hour and a half to a little over a day. They remain at their peak of brightness for a very short time. They are at minimum brightness for a comparatively prolonged period. It is extremely difficult to observe these stars. Often they can be studied only by photographic means.

Even at maximum brightness, they are so faint that we have identified only about fifteen hundred out of an estimated total of one hundred thousand in our own galaxy. As for the RR Lyrae stars in other galaxies, we shall probably never be able to discover more than an insignificant fraction. So far, astronomers have identified only a few in the neighboring galaxies.

Far better known are the Cepheid variables. Their name is derived from Delta Cephei, a star that can be made out with the naked eye. The total period of Delta Cephei from minimum to minimum is nearly five and one-half days, and its light varies from a magnitude of 3.7 to 4.4. It takes only one and one-half days for the star to reach maximum brilliance, while the descent to minimum requires four days. About fifteen hours after the downward progression begins, the star maintains the same degree of brilliance for a short time before resuming its declining course.

The average period of the Cepheid variables in our own galaxy is about seven days. Changes in magnitude vary from less than half a magnitude to somewhat more than a magnitude—a comparatively slight range.

Following are the periods and the variations in magnitude of a few representative members of the Cepheid variable group:

| Name | Period (days) | Magnitude |
|----------------|------------------|-----------|
| SU Cassiopeiae | 2.0 | 6.1–6.4 |
| Delta Cephei | 5.4 | 3.7–4.4 |
| Zeta Geminorum | 10.2 | 3.7–4.1 |
| 1 Carinae | 35.5 | 3.6–4.8 |
| SV Vulpeculae | 45.1 | 8.4–9.4 |

LIGHT CURVES

We show the variations in brightness of a star by the so-called light curve, in which we plot brightness against time. Generally we cannot make observations of a variable star throughout the cycle of its variation, if the cycle is a day or more in length, because of the alternation of day and night. Hence we have to establish the light curve of a variable star from the data we obtain on different nights. It is a tribute to the perseverance and ingenuity of astronomers that they have been able to plot so many of the light curves.

These curves show considerable variation among Cepheid variables. In some, the curve is smooth. In others, a steep ascent is followed by a gentle downward slope.

which means, of course, that they brighten more rapidly than they grow faint.

Almost all the changes in brightness in Cepheid variables are due to the changes in temperature at the surface of the stars. The surface is hottest when the star is the brightest and coolest when the star is the faintest. The coolest Cepheid variables have the longest periods.

METHOD OF DETERMINING INTERSTELLAR DISTANCE

Early in the 20th century, a new and valuable method for determining interstellar distances was discovered by Henrietta S. Leavitt, an astronomer on the staff of the Harvard College Observatory. Her discovery was the result of a study she made of the Small Magellanic Clouds, in the southern heavens. This cloud is a nearby galaxy rich in Cepheid variables. Presumably all of these stars were at about the same distance from the earth. When Leavitt had arranged these Cepheids according to their period and brightness, she found that a definite relationship existed. The brighter the star, the greater the time required to go through the period of change from one time of maximum brightness to the next. If the distance and actual luminosity of any one of the Cepheid variables could be found, it would be possible to determine the luminosities and distances of other Cepheids from their periods. This conclusion was born out by Harlow Shapley's investigations of variable stars in globular clusters. Astronomers now determined the distances of several of the brighter Cepheid variables. Then by using these results, they were able to calculate the distances of various other, generally fainter, Cepheids on the basis of the period-luminosity relationship.

The Cepheids with like periods are assumed by the period-luminosity relationship to have the same intrinsic brightness. Consequently, the difference in observed brightness shows a difference in distance from us. We know that the light that we receive from a given source at a given distance will be four times as great if we halve the distance and only one-fourth as great if

we double the distance. Let us suppose, by way of example, that a lighted candle is held at a distance of four meters from us. If the candle is then set two meters from us, we receive four times as much light. If, on the other hand, it is moved to a distance of eight meters from us, we receive only a fourth as much light. Utilizing this law of inverse relationship, we can tell that if a Cepheid variable appears to us to be sixteen times as bright as one with the same period and, therefore, the same intrinsic brightness, it must be four times nearer to us than the other Cepheid variable in question.

QUASARS AND PULSARS

Two more kinds of stars or starlike objects emit light and other kinds of radiation, such as radio waves, in varying quantities. These are quasars and pulsars. Quasars have periods ranging from one day to several months. They may be extremely distant from earth—1,000,000,000 light-years and more—yet they are very bright. A light-year is a measure of distance used in astronomy. It is the distance that light travels in one year, or approximately 9,600,000,000 kilometers. If quasars are really millions and thousands of millions of light years distant and yet are so bright, they must radiate vast amounts of energy. Astronomers who believe that quasars are this far distant can only guess about the source of all that energy. Other astronomers, however, think that quasars are nearby—astronomically speaking, that is—and therefore that their energy output is not so great.

Pulsars are stars that emit light and radio waves in very swift and regular pulses that last only a very small fraction of a second each. They seem to be far less distant than quasars—100,000 light-years.

Many astronomers believe that pulsars are very small stars, averaging only a few kilometers in diameter. They are thought to be extremely dense, with their matter consisting chiefly of subatomic particles known as neutrons. Pulsars are, in fact, sometimes known as neutron stars. They emit radiation in a tight beam, all the while rotating very fast. The overall effect has been likened to a lighthouse with a rotating beacon.

INTERSTELLAR SPACE

by Cecilia Payne-Gaposchkin

The modern astronomer is not concerned only with heavenly bodies of considerable size, such as the stars and the planets. He is also vitally interested in the gases, the solid particles of matter, and the atoms that are scattered through space between the stars. In the early 20th century this material was considered merely an obstacle to the study of the distant stars. The astronomer has come to realize, however, that interstellar ("between-star") matter is interesting in its own right. Today it is one of the greatest aids to our understanding of the structure and development of the stars and of the galaxies, or island universes, formed by associations of stars.

About 100,000,000,000 stars make up the gigantic, flattened pinwheel that is our own Milky Way galaxy, in which the sun and the planets of the solar system are located. Light, which travels some 300,000 kilometers a second, takes about 7,000 years to cross the Milky Way. This gives some idea of its immensity. Vast as it is, our galaxy is not the only one. There are others, probably hundreds of millions, in the far reaches of space. Some of them are systems like our own. Others are quite different in shape and size. Interstellar materials are found in almost all of these numberless systems of stars. We are beginning to suspect, too, that such materials also exist in the vast spaces between the different galaxies.

Our first knowledge of interstellar matter came from the bright diffuse nebulae—luminous clouds in the heavens. Perhaps the best known of these is the Great Nebula in Orion, faintly visible to the naked eye as a greenish blur. Hundreds of other bright nebulae are known. Most of them lie in or near the Milky Way. Other galaxies, especially those that are like our own, also contain bright nebulae.

THE NATURE OF NEBULAE

The early observers thought that nebulae like the one in Orion might really

consist of faint stars. As more powerful telescopes were built, astronomers tried to resolve the nebulae—that is, to break them up into stars. In certain instances, they were successful. Some of the nebulae proved to be galaxies or smaller star groups. However, a great number of them could not be resolved in this way. When astronomers viewed them even through the most powerful telescopes at their disposal, they could see only finely shredded, cloudlike structures. Late in the 19th century, it was proved, through the use of the spectroscope, that bright nebulae, such as the one in Orion, are not made up of stars.

The spectroscope and the instruments derived from it are, for the modern astronomer, as important tools as the telescope. They enable us to learn about the heavenly bodies by decoding the messages that their light brings to us. White light is made up of all the colors of the rainbow. When the light of a star or of a distant galaxy passes through the spectroscope, the light is broken up into a band of colors—a spectrum. By the position and pattern of the lines of color in this band, we can tell what elements, in the form of incandescent gas, are present in the heavenly body. Each element has its own distinguishing pattern of lines, with its own place in the long band of the spectrum.

A spectrum that shines with definitely separate, or discrete, colors is the earmark of a glowing, tenuous gas. The Great Nebula in Orion has just such a spectrum. Many other bright nebulae have spectra very like this and clearly also consist of glowing gas.

SOLID MATERIAL

Not all the bright nebulae have gaseous spectra, however. The spectrum of the nebulous material that surrounds the Pleiades, a cluster of rather hot stars, resembles those of the bright stars within it. But we should be wrong in concluding that this nebula itself consists of stars. Actually,



Bart Bok

Mt. Wilson and Palomar Observatories

Interstellar space may hold the answer to many questions about the origin and nature of the universe. Is it the material from which stars are made? Does it contain the building blocks of life? Astronomers are seeking answers to these puzzles as they focus their telescopes on various nebulae, including that near IC 2944 (above) and that near NGC 2264 (right).





The Great Nebula, M42, in Orion. It appears to the naked eye as a faint blur but it is really a churning mass of glowing atoms, shining by the light of a group of stars near it.

Yerkes Observatory

it is simply reflecting the light of the bright stars near it. We assume that it must contain solid particles. Otherwise, it would not be able to reflect starlight as well as it does. It is probable that the Pleiades nebula contains not only solid particles but also gases.

We have another way of knowing that solid matter is present in interstellar space. Clouds of particles screen off, or obscure, the light of more distant stars. Some clouds, like the Coal Sack in the southern sky, are opaque and have well-defined edges. Others can be located by the general dimming of the stars behind them and, what is more important, by the reddening of the starlight that passes through them. If the observed color of a star is redder than we know it should be, we can be sure that it is screened by a cloud that is absorbing some of the starlight. We can also tell how much

of this obscuring material there is. Most of this matter lies in the plane of the Milky Way.

Solid particles are not the only obscuring bodies in the space between the stars. Individual atoms, too, can obscure starlight. Each atom absorbs radiation of the same wavelength as the atom itself emits. The radio telescope is a spectacular tool for discovering the existence of atoms in interstellar space. It detects electromagnetic radiation in the radio frequencies. The wavelength of red light (also a form of electromagnetic radiation), to which the eye is sensitive, is about 0.000065 millimeters. The radio telescope can detect radiation with a wavelength of about 200 millimeters. It happens that the atom of hydrogen, commonest of all substances, emits radiation of about this wavelength. A Dutch

scientist brilliantly predicted that radiation from hydrogen should exist in the sky. American physicists built a radio receiver to search for it, and signals sent out by interstellar hydrogen were picked up. The detection of radiation emitted by hydrogen marked a new era in the history of astronomy.

GLOWING GASES

The bright nebulae, as we have seen, lie near the galactic plane. A fine gaseous haze, much fainter than these conspicuous knots of gas, pervades the Milky Way. Powerful instruments are required to photograph its spectrum, for it is six thousand times fainter than the Great Nebula in Orion. This haze contains about one atom per cubic centimeter, on the average. The enormous majority are atoms of hydrogen. They shine because of the radiation that falls on them—the very faint light of many distant stars. There is evidence that atoms

of oxygen, nitrogen, and sulfur are also present.

The bright gaseous nebulae are found only near stars whose light corresponds to a temperature greater than about 13,900° Celsius at the surface. The light of these nebulae shows the spectral colors (colors of the spectrum) characteristic of glowing hydrogen and also certain colors associated with helium, carbon, nitrogen, oxygen, sulfur, and other atoms. The Orion nebula contains over three hundred atoms per cubic centimeter. Here, again, most of the material is hydrogen.

What makes these atoms glow? They are close to a hot star. The hydrogen and helium atoms absorb extreme ultraviolet light from the star and give it out again in the form of their particular radiations, or wavelengths. The oxygen, nitrogen, and sulfur atoms give out light after they collide with electrons that have been detached from other atoms—mainly hydrogen

The Milky Way is pervaded by a fine gaseous haze, composed mainly of hydrogen atoms. These atoms glow because radiation from distant stars falls on them.

Yerkes Observatory



atoms—by the high-temperature radiation of the star. If the material had densities comparable to those in stellar atmospheres, collisions with other atoms would, in a short time, rob these atoms of the energy conferred by the electrons. But the density of the nebulae—several hundred atoms per cubic centimeter—is so low that collisions may not take place for months. Under these conditions, the atoms of oxygen, nitrogen, and sulfur will radiate the lines seen in the nebular spectrum, which are called the forbidden lines. They are “forbidden” only in the sense that they are suppressed under stellar or laboratory conditions by constant collisions with neighboring atoms.

The bright nebulae that surround stars with temperatures less than 13,900° Celsius differ from nebulae such as the one in Orion. The nearby star gives out too little ultraviolet light to excite the bright spectral lines. Here the major effect is the reflection of the starlight by solid particles. Though atoms will certainly be present, too, the star is not hot enough to give them the energy needed to make them glow.

OUTLINING GALAXIES

Interstellar hydrogen and solid particles help us to make out the details of some of the galaxies. These are divided into several types. In the spiral galaxies, huge, curving arms are set around a nucleus, which may be spherical or in the form of a straight bar. Our Milky Way is a good example of a spiral galaxy. Other galaxies, called irregular, have no marked symmetrical form. Still others, the elliptical galaxies, are distinguished only by their elliptical shape.

When astronomers observed the interstellar hydrogen in the faint nebulosity of the Milky Way with the radio telescope, they found that the interstellar hydrogen gas outlines the spiral structure of our stellar system. The gas lies most densely in the spiral arms and thins out between them. The hydrogen is accompanied by absorbing particles, which follow the course of the spiral arms. In the galaxy called the Great Spiral in Andromeda, the spiral arms are found to be studded with bright gaseous

nebulosities and associated with well-marked lanes of obscuring materials.

The chaotic, irregular galaxies are even richer in bright and dark nebulous material than the spirals. Many even of the featureless elliptical systems, which contain few if any solid particles, are pervaded by a very diffuse haze of glowing atoms.

Astronomers have found that several distinct clouds, moving with different speeds, lie between us and many distant stars. Evidence has shown that these separate clouds are associated with different spiral arms of our galaxy. Each is moving with its own speed and is silhouetted against the light of very distant stars.

The interstellar gases furnish evidence about conditions in space between the stars. In the region of faint, hazy nebulosity, the temperature of a solid particle would be a few degrees above absolute zero, which is -273° Celsius. The starlight falling upon such particles is so faint that they are intensely cold.

Like the glowing atoms, the solid interstellar material tends to lie in the plane of our galaxy. From what we see of other spiral systems, however, we may suppose that it is confined to the outer regions, and absent from the center of our system.

SMALL CRYSTALS

Interstellar solids must be very finely divided to absorb as much starlight as they do. From studies of the motions of stars, we know that these solids cannot have a density greater than 0.0000000000000000000000008 grams per cubic centimeter. At a density as low as this, the particles may have diameters of about 0.000015 centimeters. However, the size required to account for the observed absorption depends on whether the particles are metallic or not. If they are metallic, they need not be so small. Metallic atoms do not seem to predominate in the universe. In the sun, which is a fairly typical example of cosmic material, there is less than one metallic atom in six thousand.

The extremely high reflecting power—greater than that of snow—of nebulae suggests that most of the solids in space are



Both, Mount Wilson and Palomar Observatories



Interstellar hydrogen and solid particles have helped to determine the outlines of many galaxies. In many spiral galaxies, such as NGC 4826 (above) and NGC 4565 (right) in Coma Berenices, solid interstellar material tends to lie in the plane of the galaxy.

tiny frozen crystals of compounds of the commonest chemical elements. Perhaps they are substances such as ammonia, methane, and even water, built of the cosmically common atoms of hydrogen, carbon, nitrogen, and oxygen.

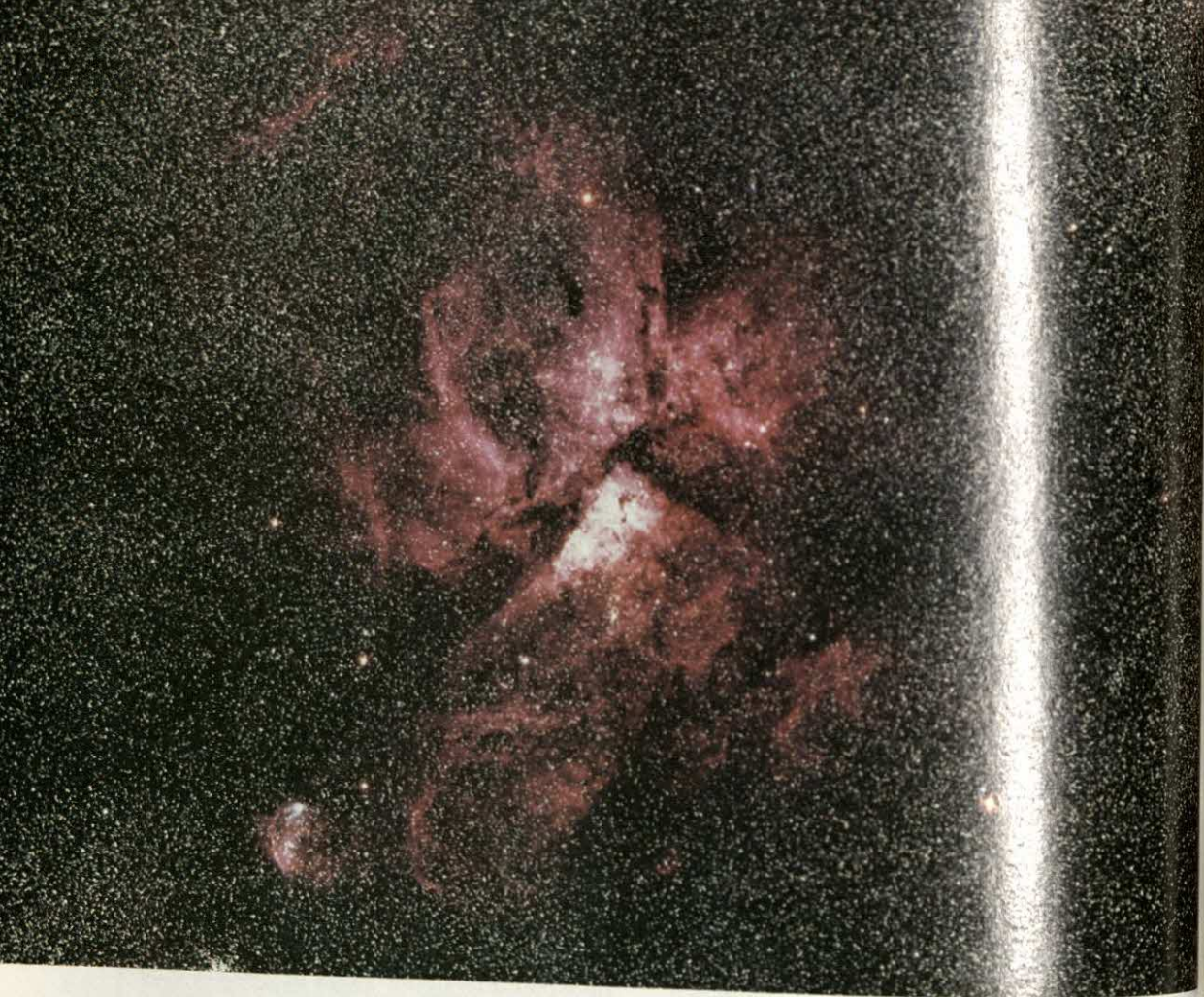
Other evidence indicates that the interstellar particles must be elongated crystals, for the light of the distant stars is polarized. This means that the light vibrations tend to be lined up perpendicularly to the light ray. The polarization is associated with interstellar reddening. Spherical particles would not have this effect. Moreover, the particles seem to be lined up more or less parallel in any one region.

There is much controversy as to how these observations should be interpreted. The consensus is that the elongated particles are lined up by magnetic fields in interstellar space. One theory suggests that small highly magnetic particles are embedded in the icy grains.

Interstellar matter does not serve merely as a space filler, but rather is impor-

tant in the formation of stars and galaxies. It is believed that, under the right conditions, an interstellar cloud of gas and dust will condense and assume a spherical shape. This sphere is a protostar, or early stage, which will eventually lead to the birth of a new star.

Radio telescopes have detected a variety of chemical elements and compounds in space. These substances exist as individual atoms, as molecules, or parts of molecules. Many of these molecules are biological in nature. That is, they are chemically identical to substances found in or produced by living things on earth. Among them are hydrogen, water, ammonia, and carbon-containing compounds such as formaldehyde and cyanogen. A number of scientists believe that these interstellar materials may be the building blocks from which life as we know it could develop elsewhere in the universe. Thus the study of interstellar space, particularly by means of the radio telescope, is extending the science of astronomy into the realm of biology.



The Eta Carinae Nebula, a glowing beauty in the Milky Way.

Kitt Peak National Observatory

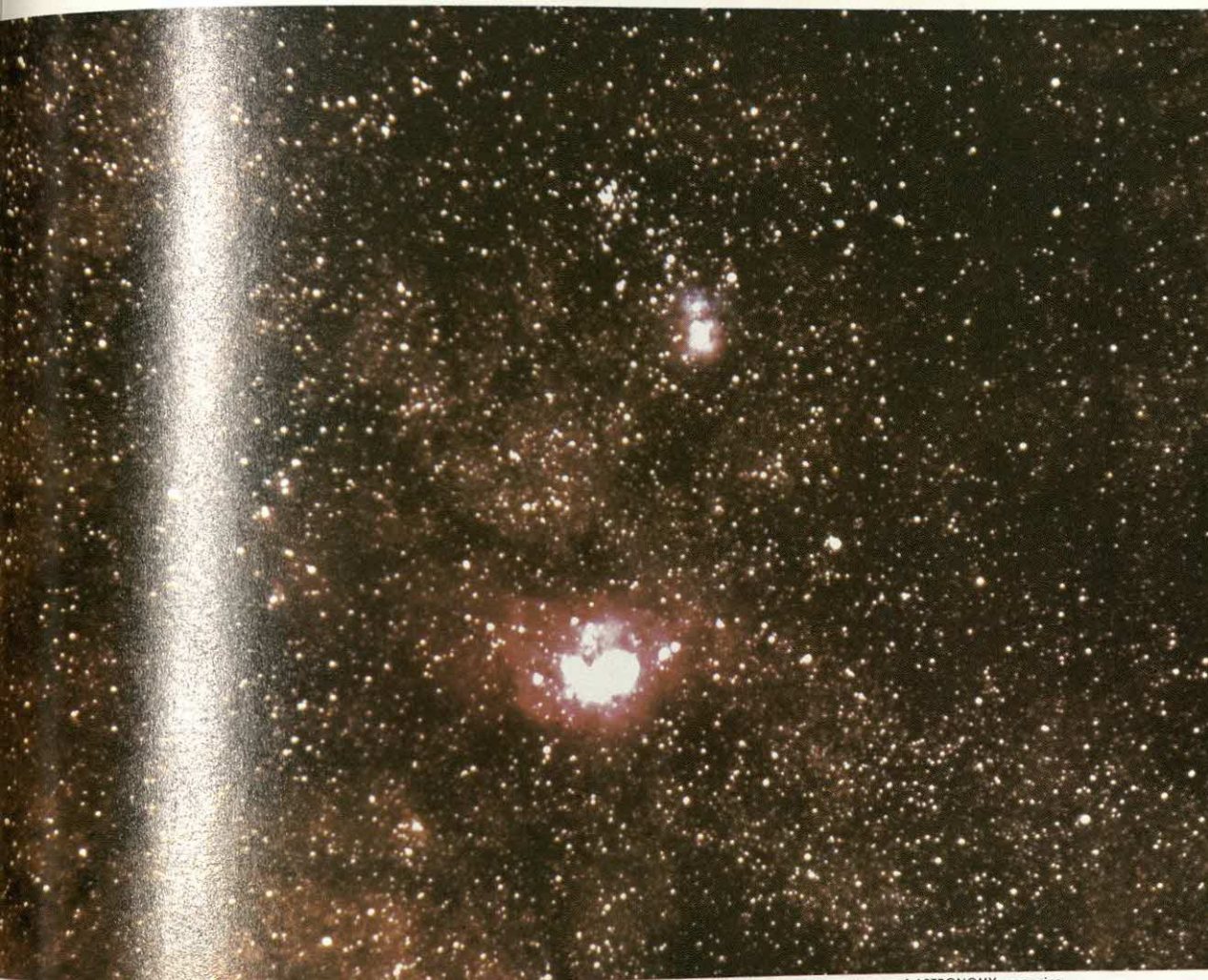
THE MILKY WAY

The Milky Way is one of the most striking sights in the night skies. It is too faint to be seen in bright moonlight or amid the myriad lights of our large cities, but on moonless nights in the country we can easily make out the outlines of its cloudy track of light across the heavens. If we peer at it through a powerful telescope, we realize that the Milky Way represents the combined light of vast numbers of stars, which cannot possibly be made out individually without a telescope.

We know today that the Milky Way forms part of a vast system of stars, to which our sun belongs. In the past, how-

ever, the Milky Way was a celestial puzzle. It was explained in various ways in Greek and Roman mythology. Some writers called it the highway of the gods, leading to their abode on Mount Olympus. Others held that it sprang from the ears of corn dropped by the goddess Isis as she fled from a pursuer. Still others believed that the Milky Way marked the original course of the sun god as he sped across the skies in his chariot.

In medieval times pilgrims associated the Milky Way with their journeys to various sanctuaries. In Germany, for example, it became known as Jakobsstrasse, or James' Road, leading to the shrine of St.



Dennis Craig Stayer, courtesy of *ASTRONOMY* magazine.

The Lagoon and Trifid Nebulae, part of the constellation Sagittarius.

James at what is now Santiago de Compostela, in Spain. In England it was called the Walsingham Way. It was associated with the pilgrimages to the famous shrine of Walsingham Abbey. The pilgrims of those days did not seriously believe that the Milky Way had anything to do with their travels. Rather, they saw it lying overhead, a misty path in the heavens and their belief in the universal kinship of all things caused them to find comfort in its presence.

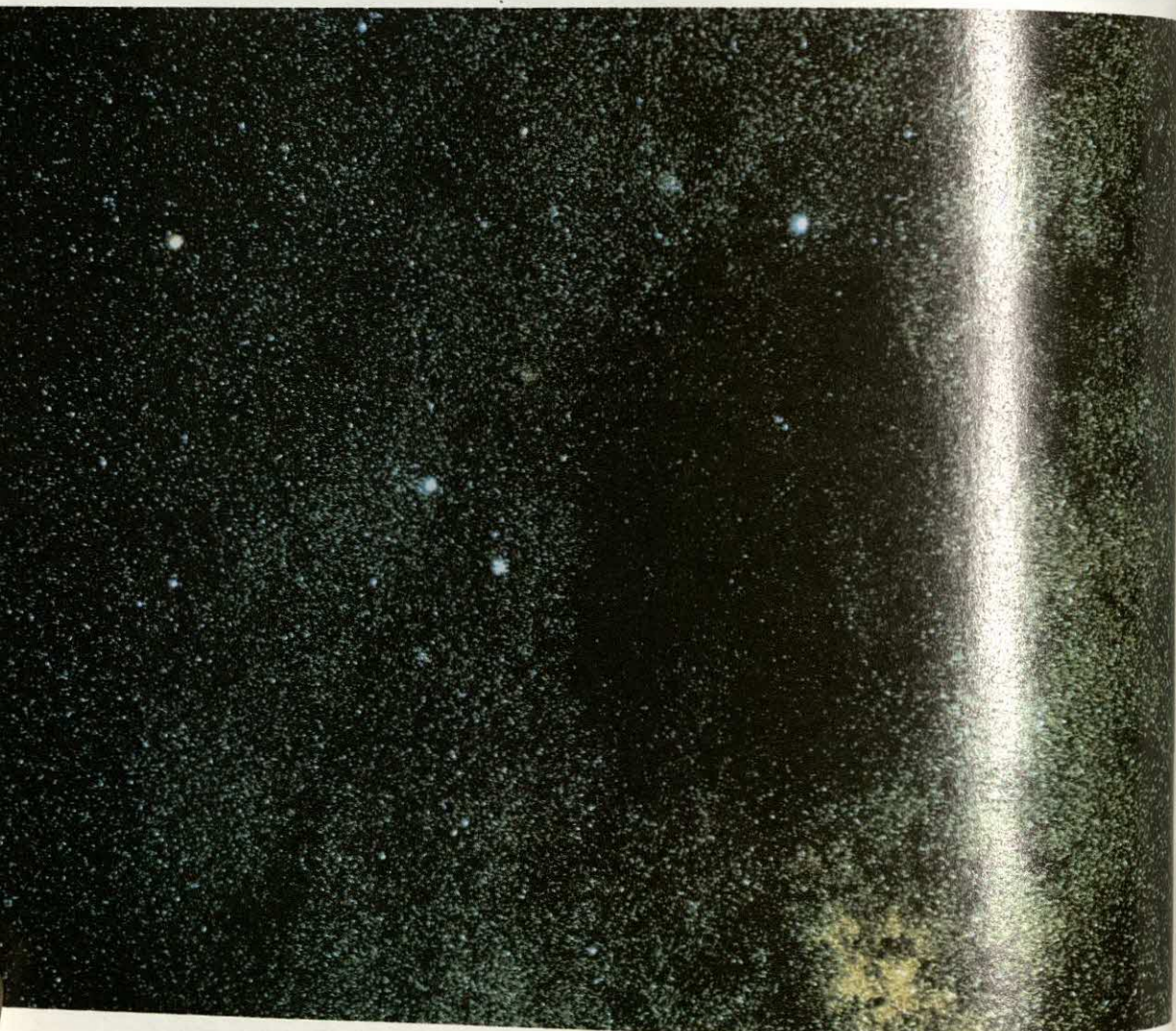
The best time to see the Milky Way is on an autumn or winter evening (in the country, as we pointed out). It is then highest in the heavens and therefore its light is least affected by our atmosphere. It is seen to stretch like a vast, ragged semicircle over the skies of the Northern or the Southern Hemisphere. Actually, it traces a rough cir-

cle, for it is continued in the other hemisphere.

IRREGULAR SHAPE

The path traced by the Milky Way is full of irregularities. It is by no means a simple stream of stars. Its average width is about twenty degrees, but it varies considerably, both in width and in brightness. Even with the naked eye, one can make out something of its irregular detail when the atmosphere is unusually clear and there is no moon. When viewed under such conditions through a good telescope, the Milky Way is a truly exciting spectacle.

Its general effect has been likened to that of an old, gnarled tree trunk, marked here and there with prominent knots. As details become clearer in a telescopic view,



The southern part of the Milky Way, showing the dark nebula called the Coal Sack. This "inky-black hole" is really a mass of obscuring matter, surrounded by bright stars.

Ronald Royer, courtesy of ASTRONOMY magazine.

we see that at one point the Milky Way may consist of separate stars scattered irregularly upon a dark background. Elsewhere, there are gorgeous star clusters. In many places, the track is engulfed in nebulosity, in which a great many stars are embedded.

A powerful telescope reveals a great many dark bands in the Milky Way. Sometimes these dark bands are parallel; sometimes they radiate like branches from a common point; sometimes they are lined

with bright stars. In certain places, they are quite black, as if utterly void of content. In others, they are slightly luminous, as if powdered with small stars.

THE COAL SACK AND OTHER DARK AREAS

Large, dark areas occur here and there. The most famous of these is the so-called Coal Sack, which is near the constellation called the Southern Cross. Just before the

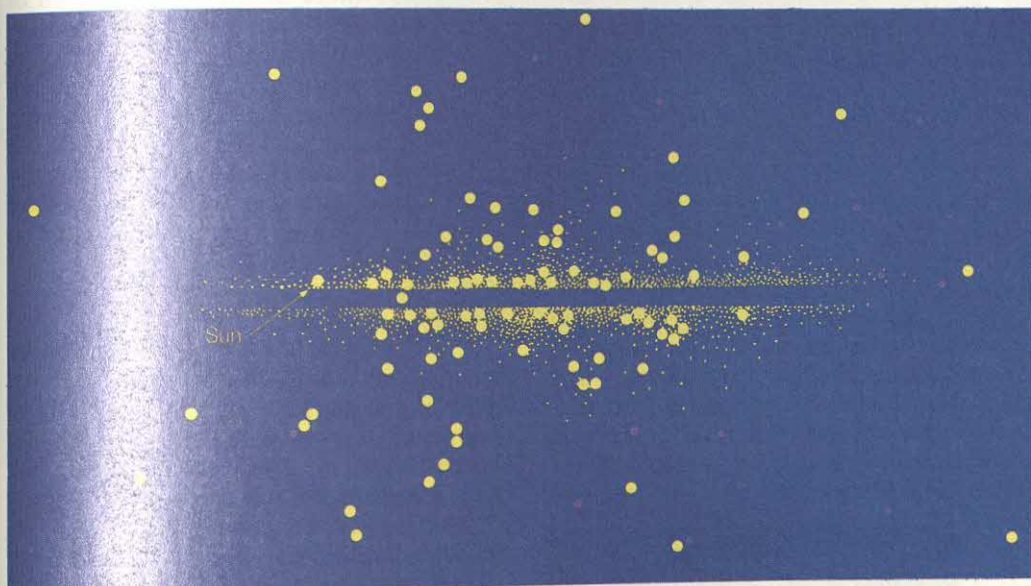


Diagram of the Milky Way, edge on, showing stars and clusters. The sun is about two thirds of the way from the center.

Milky Way divides into two branches in the southern constellation Centaurus, it broadens. It now becomes studded with a collection of brilliant stars, so that this is one of the most resplendent areas in its whole course. Right in the center of this host of bright stars, near the four stars that form the Southern Cross, is the inky black cavern of the Coal Sack.

The Coal Sack is by no means unique. There are many similar black areas in the Milky Way, though they are generally less clearly defined and less striking in appearance. The American astronomer Edward E. Barnard described one of these, in the constellation Sagittarius, as "a most remarkable, small, inky-black hole in a crowded part of the Milky Way, about two minutes in diameter, slightly triangular, with a bright orange star on its north north-westerly border, and a beautiful little star cluster following."

STUDYING THE SYSTEM

As we pointed out, the starry band that extends across the heavens is really part of

a galaxy, or system of stars. Every star that can be seen with the naked eye belongs to this vast system, including our sun and the other members of our solar system.

Various methods have been used to solve the mystery of its structure. The 18th-century German-English astronomer William Herschel decided to attack the problem by making a survey of the stars. He called his method "star gauging." It consisted of counting all the stars visible in a reflecting telescope that had a 45-centimeter mirror and a field 38 centimeters in diameter.

The Dutch astronomer J. C. Kapteyn developed an even more elaborate method for determining the distribution of stars in the heavens. He selected 206 areas distributed uniformly over the whole sky, and he urged astronomers to determine the apparent magnitudes and other data for all the stars in these regions. A number of the great observatories of the world took part in the cooperative venture suggested by Kapteyn. Two outstanding contributions to the program are the Mount Wilson Cata-

logue and the Bergedorfer Spectral-Durchmusterung (Bergedorf Spectral Catalogue). The latter gives not only the apparent magnitudes of the stars but also the spectral classes of the stars.

Another way of determining the dimensions and structure of our galaxy is to observe the positions and distances of some of its members, such as the Cepheid variables. We can find the approximate distances of these star groups from the earth if we know their apparent magnitudes.

The chief difficulty in determining the extent and structure of our galactic system is that we are embedded in it so deeply that we cannot obtain a comprehensive idea of it through mere observation. It is almost as if we were required to make a map of New

Both photos: © California Institute of Technology and
Carnegie Institution of Washington



York City, for example, from a vantage point somewhere in a crowded section of the Bronx. However, we do have a clear over-all view of a number of other galaxies, such as the magnificent spiral galaxy in Andromeda (Messier 31) and the Whirlpool in Canes Venatici (Messier 51). They enable us to draw a number of plausible conclusions about our own galaxy. The radio telescope has also provided information about our galactic system.

A SPIRAL GALAXY

An analysis of our galactic system by the methods described above indicates that most of the stars it contains are crowded into a sort of wheel with a pronounced hub. When viewed edge on, the wheel and hub look something like the illustration on page 209. Of course, since the sun, our own star, is located within the wheel, we cannot see the wheel edge on. We infer its shape from viewing other galaxies whose form we can make out and think are similar. Our sun does not lie anywhere near the center, or hub, of the wheel. It is at a distance of about two thirds of the way from the center to the outer rim. The wheel as a whole is inclined at an angle of 62° to the plane of the celestial equator.

The stars that make up the wheel show a distinct spiral pattern. Astronomers therefore classify our galaxy as a spiral galaxy. The spiral arms spring from a nucleus or core at the center of the galactic system. In the spirals, we find individual bright stars, star clusters, bright nebulae, and a great deal of obscuring matter. This obscuring matter is made up of dust particles and various gases. The general haze it causes has been detected through the dimming and reddening effects that it produces.

DARK NEBULAE

The name "dark nebulae" has been given to the masses of obscuring matter which have no stars nearby to illuminate

Planetary nebulae are huge clouds of luminous gas surrounding a very hot star. They are usually distinct and symmetrical. At left, two planetary nebulae: the Ring nebula in Lyra (top) and the Dumbbell nebula in Vulpecula (bottom).

The Pleiades—a galactic, or open, cluster of stars. The gas and dust nearby scatter the starlight to form diffuse nebulae around the stars.



© California Institute of Technology and Carnegie Institution of Washington

them. Some of them can be seen quite clearly with the naked eye and have been known to astronomers for a long time. The earlier astronomers thought of them as gaps in the starry firmament. The dark nebulae, or masses of obscuring matter, cut off our view of vast numbers of stars that lie beyond them. To obtain an approximate idea of the form of the galactic system, certain observers have counted the stars in various parts of the sky and have then made allowances for the obscuring matter and estimated the form of the galaxy.

Obscuring matter is responsible for the large dark areas such as the Coal Sack. At one time, the attempt was made to explain

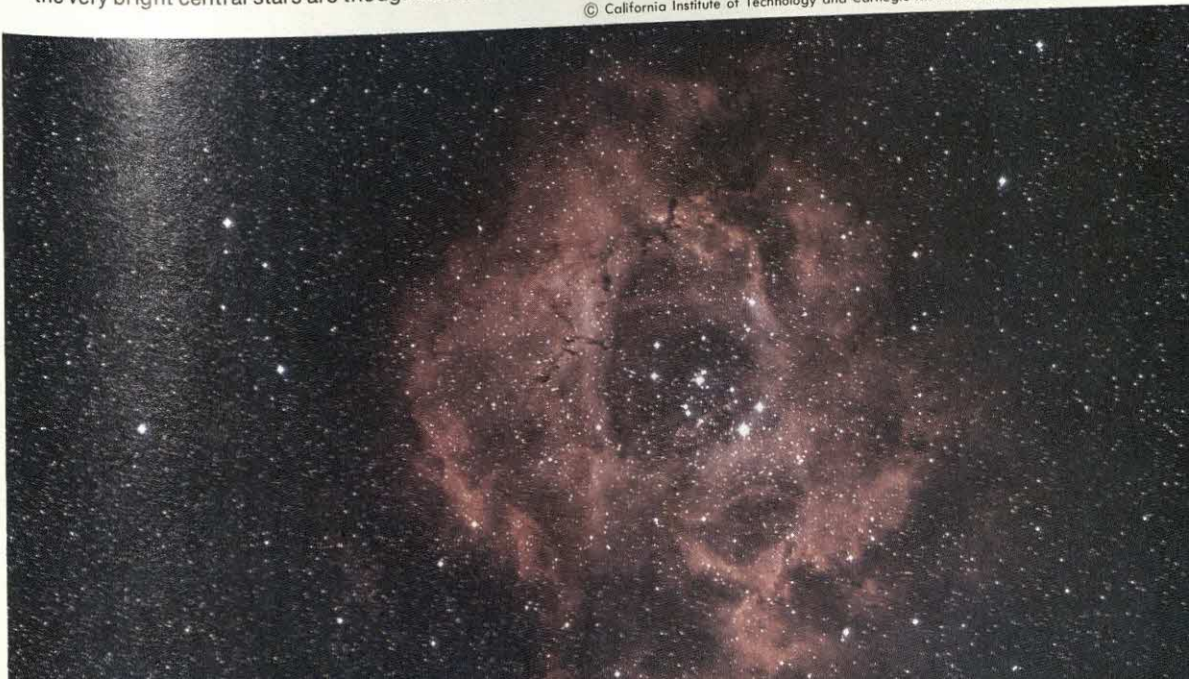
the Sack as an optical illusion. This hypothesis never seemed too convincing, however. One could not explain away the sharp distinctness of the outline of the Coal Sack, its huge size, its utter darkness, and the even brightness of the starry edge surrounding it. The existence of obscuring matter, now fully proved, offers a simple explanation of what was once a mystery.

STAR CLUSTERS

The clouds and wisps of bright nebulosity that we mentioned before are the bright nebulae, whose atoms have been excited by the hot stars in the vicinity or else reflect the light of nearby stars.

The Rosette Nebula in Monoceros. The globules of luminous gas surrounding the very bright central stars are thought to be contracting to form new stars.

© California Institute of Technology and Carnegie Institution of Washington





© California Institute of Technology and Carnegie Institution of Washington



© California Institute of Technology and Carnegie Institution of Washington

It is interesting to note that the most luminous stars are all to be found in the wheel of our galaxy. The so-called galactic clusters are also confined to this area. They are sometimes called "open clusters." They consist of groups of hot stars, each group consisting of several hundred stars. They excite the atoms of dust or gases in their vicinity; hence they are embedded in bright nebulae. Among the best-known of the galactic clusters are the Pleiades, the Hyades, and Coma Berenices.

Left, top: the Crab Nebula in Taurus—the remnant of a violent supernova explosion that occurred in 1054 A.D. Left, bottom: the Trifid Nebula, a cloud of glowing gas and dark lanes of dust, in the constellation Sagittarius.

The globular clusters are distributed through a roughly spherical region bisected by the plane of the galaxy. The American astronomer Harlow Shapley held that the main aggregation of stars in the galactic system is arranged in the form of a wheel and that this is enclosed in a roughly globular haze of stars—globular clusters and others. Globular clusters differ in many respects from the galactic clusters. For one thing, each one contains hundreds of thousands of stars, or perhaps even millions, compactly and symmetrically grouped. They are comparatively free of gas or dust.

GALACTIC CORE

The core or nucleus of the galactic system is so heavily obscured by clouds of dust that astronomers have a very imperfect idea of its structure. It has never been successfully observed photographically. With the aid of the radio telescope, however, we have been able to determine the existence of a heavy concentration of stars in this area. It has been estimated that these stars account for something like one half of the total mass of our galactic system.

THE ROTATION OF OUR GALAXY

The entire galactic system is rotating around an axis which is at right angles to the wheel of stars. Outside the dense nucleus of the galaxy, the speed of rotation decreases and the period increases with greater distance from the center. Corresponding differences in period and speed can also be noted in the planets that revolve about the center of the solar system—that is, the sun. The nearer a planet is to the sun, the faster it revolves around it and the shorter the period of revolution. For example, in the case of Mercury, the planet nearest the sun, the period is only 88 days, compared with the period of 248.43 years of the planet Pluto, whose orbit is farthest from the sun.

Obviously, then, the stars between the sun and the galactic center go around the

center of the galaxy more rapidly than the sun does and eventually overtake and pass it. On the other hand, the stars between the sun and the outer rim of the wheel revolve around the center of the galactic system more slowly than the sun does. As a result, they lag farther and farther behind our star in their voyaging.

THE DIMENSIONS OF OUR GALAXY

In the foregoing pages, we have given you some idea of the form of our galaxy. Astronomers have also sought to discover its dimensions in space. Recent researches, based on the study of special types of variable stars (the Cepheid variables) and on the visual, photographic, and spectrographic analysis of globular clusters have revealed how extraordinarily great these dimensions are. This is natural enough when we consider that our own solar system, vast and complex though it is, is an infinitesimally small part of the galactic system. The estimates of the dimensions of our galaxy vary considerably, but even the smallest are almost overpowering when we seek to grasp their significance.

According to one of the "small" estimates, made by the American astronomer Heber D. Curtis, a pulse of light, starting from one edge of the galactic system and traveling at the speed of over 300,000 kilometers a second, would take from 20,000 to 30,000 years to reach the other edge. According to Harlow Shapley, Curtis's estimate, staggering though it may seem to us, was far too modest. If we are to accept Shapley's figures, we must tax our imagination still further. He maintains that it would take about 100,000 years for a pulse of light to travel from one confine of the galaxy to the other. In other words, the diameter of the galactic disk is approximately 100,000 light-years.

We must remember that our galaxy, of which the solar system forms a part, is only one of millions of other isolated galactic systems. Each of these island universes contains millions upon millions of stars. The nearest—two irregular ones, called the Magellanic Clouds—are about 100,000 light-years distant from us.

Beyond that distance, the heavens are studded with galaxies. When we reach the

A nearby planetary nebula—NGC 7293, found in the constellation Aquarius.

Société Astronomique de Suisse





Galactic cluster NGC 457 in the constellation Cassiopeia. In this type of cluster, the association of stars is not as complex and condensed as in a globular cluster

Société Astronomique de Suisse

limit of vision—two thousand million light-years—with our greatest telescope, the 500-centimeter giant of Palomar Mountain, the island universes show no signs of thinning out. It is clear that this telescope, gigantic though it is, has not yet reached the extreme “edge” of the universe. With improvements in instrumental and photographic technique, it may be possible to extend our horizons still farther.

THE “EDGE” OF THE UNIVERSE

Whether we shall ever approach the “limit” of the universe is extremely doubtful in view of a certain sobering consideration. If the universe is 10,000,000,000 years old, as some scientists believe, our horizon will be limited to a distance of 10,000,000,000 light-years from the earth, even if the universe actually extends far beyond that distance. The reason is that the light from galaxies beyond that mark will not have been able to reach us since the beginning of time. The same is true of the radio waves emanating from such galaxies, since light waves and radio waves travel at the same speed—roughly, 300,000 kilometers

per second. Such galaxies, therefore, will remain a deep mystery to astronomers, unless some hitherto unsuspected method of detecting them is discovered.

This nebulosity around *Nova Persei* is expanding. Is the entire universe still expanding or shall we eventually see signs of an “edge” to our universe?

Mt. Wilson and Palomar Observatories



GALAXIES

by Gerard de Vaucouleurs

Galaxies are vast systems of stars. Our own star, the sun, is a part of the galaxy known as the Milky Way. Galaxies populate the depths of space out to the limit of penetration of the largest telescopes. The total number that could be photographed with the 500-centimeter telescope on Palomar Mountain may come to a thousand million or so. It is reasonable to assume that many more could be brought into view with an even more powerful telescope.

EARLY IDEAS ABOUT GALAXIES

The idea that stellar systems existed outside our Milky Way galaxy had been discussed by philosophers and astronomers as early as the eighteenth century. In the second half of the century, the great German philosopher Immanuel Kant advanced the theory that the celestial objects observed as faint, hazy patches of light and called nebulae were really systems of stars. Kant's theory was confirmed toward the end of the eighteenth century by the British astronomer William Herschel. With the aid of large reflecting telescopes that he himself built, Herschel discovered thousands of new nebulae. In the early nineteenth century, the German scientist Alexander von Humboldt gave the nebulae the picturesque name of "island universes." Few astronomers of that period doubted that the nebulae were really made up of stars. It was believed that with the use of more powerful telescopes it would be possible to make out the individual stars in an island universe.

The problem of the nature of the nebulae was reopened in 1864. In that year, the British amateur astronomer William Huggins observed that many of the irregular, diffuse nebulae, such as the one in the constellation Orion, were composed, not of

Top: the Andromeda galaxy, a large spiral galaxy that is a member of a local cluster that also includes our Milky Way system. Middle: galaxy M82, an extremely active galaxy. Bottom: photo showing typical spiral galaxy with arms spiralling around a central core known as the nucleus.

California Institute of Technology



California Institute of Technology and Carnegie Institution of Washington

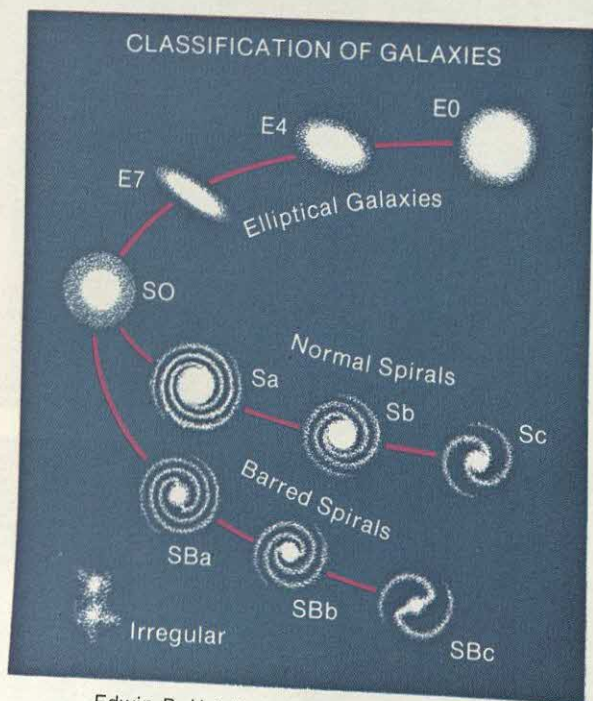


California Institute of Technology and Carnegie Institution of Washington





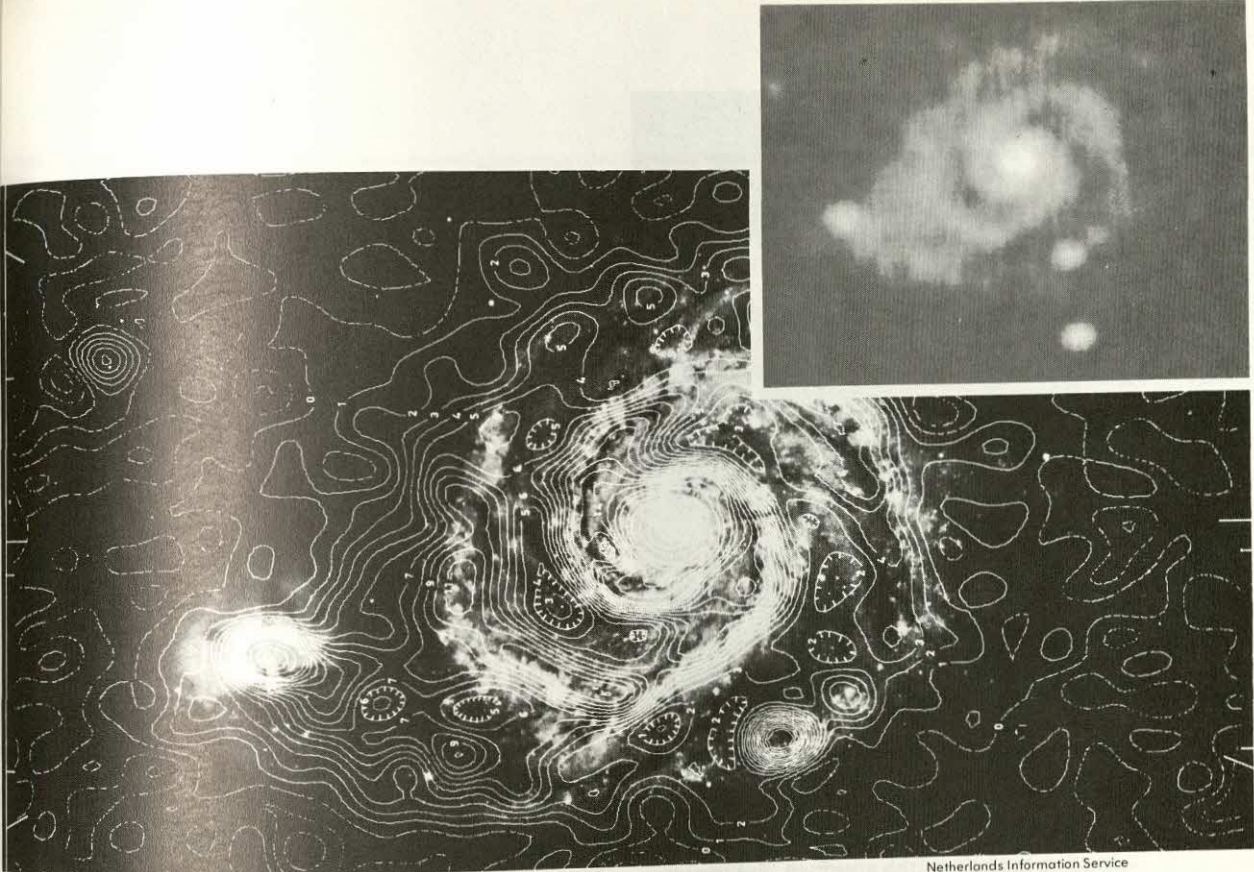
Lick Observatory



Edwin P. Hubble's system, summarized above, for the classification of galaxies into four main types has been accepted with some revisions by astronomers the world over.

stars, but of extremely rarefied gas. The spectrum of this gas consisted of isolated bright lines, of which a pair of green lines were the most conspicuous. On the other hand, other nebulae, like the one in the constellation Andromeda, were whitish in color, and their spectra appeared more or less continuous, like that of our own star, the sun. Apparently, then, there were two kinds of nebulae—one kind made up of gases, the other of stars.

It was not until the first quarter of the twentieth century that the full meaning of the distinction between the gaseous nebulae and the stellar nebulae was firmly established. The existence of "extragalactic nebulae" (galaxies outside the Milky Way galaxy) was then placed beyond doubt. This was due to the work of the American astronomers Heber D. Curtis and Edwin P. Hubble—particularly Hubble. In the years between 1924 and 1936, Hubble pushed the exploration of extragalactic space from the nearer galaxies, at distances of the order of a million light-years, to the limit of penetration of the 250-centimeter telescope at Mount Wilson Observatory, at distances a thou-



Netherlands Information Service

Three views of the Whirlpool galaxy. Left: a conventional photo of the galaxy. Above right corner: a radio "photograph," produced by processing radio data to obtain a visual image. Above: a radio-wave "map" superimposed on a conventional photo. The closer the map lines, the more intense the radio waves.

sand times greater. (A light year is a unit of distance used in astronomy. It is equal to the distance that light travels in one year, or approximately 9,500,000,000,000 kilometers.) The more distant galaxies that have been photographed with the 500-centimeter telescope on Mount Palomar in California may be at distances of several thousand million light-years.

CATALOGUES AND GALAXY COUNTS

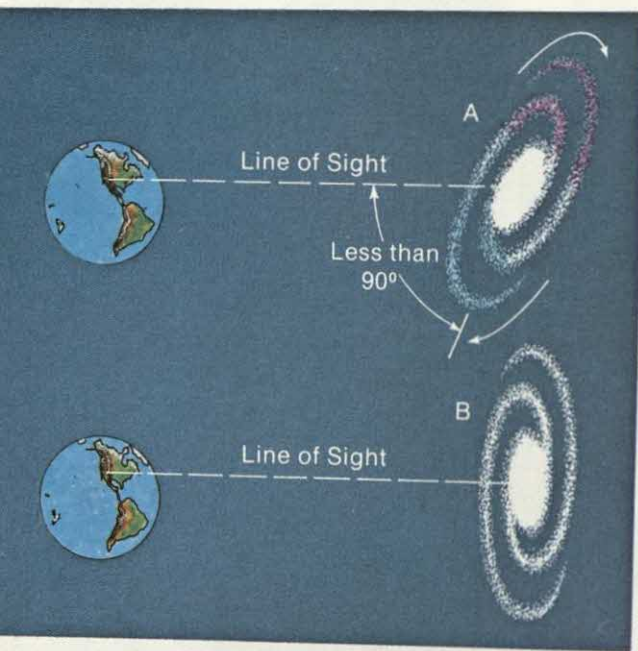
The first catalogue of nebulae and star clusters was published in 1782 by the French astronomer Charles Messier. The numbers in his catalogue are still in use for the brightest objects that he listed. For example, the Andromeda nebula—which we now know to be a galaxy—is often referred to as Messier 31, or M 31.

In the year 1783, William Herschel, in England, began a series of systematic surveys of the northern skies. His work was continued and extended to the Southern

Hemisphere by his son, John Herschel, who published in 1864 a *General Catalogue* of 5,000 nebulae and clusters. In 1888, John L. E. Dreyer, of Armagh Observatory, Ireland, published a *New General Catalogue* (NGC), which included 7,814 objects. Two supplements, issued in 1895 and 1908, brought the total of catalogued nebulae and star clusters to over 13,000. The NGC numbers are in universal use; for instance, the Andromeda nebula is NGC 224, as well as M 31.

After the introduction of photography in astronomy, the number of recorded nebulae increased rapidly. Photographic surveys and counts of all galaxies of a certain brightness began to be compiled. One of the first, published in 1932 at the Harvard College Observatory by American astronomers Harlow Shapley and Adelaide Ames, included 1,249 galaxies brighter than the thirteenth magnitude.

Magnitude is a measure of the appar-



Suppose a galaxy is inclined to the observer's line of sight at an angle of less than 90 degrees, as in A above. Then one side of the galaxy would be moving away from the observer, and the lines of a spectrum taken of this section would be displaced toward the red part of the spectrum. But the other side of the galaxy would be moving toward the observer, and the lines would be displaced toward the blue part of the spectrum. If a galaxy is seen face-on, as in B, its rotation cannot be determined by an examination of the lines of its spectrum.

ent brightness of a star. In early catalogues, first magnitude was assigned to the brightest stars; sixth magnitude to those just visible with the naked eye. Later as powerful telescopes revealed stars never seen before, the scale of magnitude was extended. The faintest stars that can now be observed are of about magnitude 23. A few stars brighter than those assigned first magnitude have also been found. These very bright stars are given negative magnitudes. The brightest star in the sky is the sun, given a magnitude of -26.72 .

Exhaustive counts of galaxies brighter than the eighteenth magnitude have been made by the American astronomer Charles D. Shane and his collaborators at the Lick Observatory, in California. In 1936, Hubble published a series of sampling counts to the twentieth magnitude in small regions regularly distributed over the celestial

sphere. From such counts as these, it has been estimated that there are about 1,000,000 galaxies brighter than magnitude 18 and over 50,000,000 brighter than magnitude 21.

The different catalogues and counts that have been prepared by astronomers from Messier's time to the present have provided the material for studies of the distribution of the galaxies.

APPEARANCE AND CLASSIFICATION

As we have pointed out, the extragalactic nebulae appeared in the small telescopes of the early observers as faint, diffuse patches of light. They seemed to be either circular or elliptical. Internal structure was first noted when larger telescopes became available. In 1845, Lord Rosse (William Parsons) and his assistants, at Parsonstown in Ireland, discovered that certain nebulae were spiral in form.

In 1925, Hubble proposed a galaxy classification that has been accepted, with certain modifications, by astronomers the world over. In its original form the classification divided galaxies into four main classes, as follows:

(1) *The ellipticals (E)*. They have a smooth structure, from a bright center out to vaguely defined edges.

(2) *The normal spirals (S)*. They show spiral arms or whorls emerging from a bright nucleus.

(3) *The barred spirals (SB)*. Their spiral arms emerge at the extremities of a bar across the nucleus.

(4) *The irregular galaxies (I)*. Some of these are of the same type as the two galaxies called the Magellanic Clouds and are classified as magellanic irregulars (Im). Others are so chaotic in appearance that they are simply listed as irregulars (I). Hubble distinguished three stages among both normal and barred spirals, labeling them a, b, and c. The relative size of the nucleus decreases from a to c; the development of the arms increases from a to c.

The Hubble classification has been revised in order to include certain new types and subtypes. The classification now includes the lenticular type (SO), having

the form of a double-convex lens. This class combines the smooth structure of the ellipticals with the luminosity distribution of the spirals. New stages d and m have been added to the spirals. These stages form the transition between stage c and the magellanic irregulars (Im).

STAR POPULATIONS

The German-born astronomer Walter Baade distinguished two basic types of star populations in galaxies, as follows.

Type I is found in irregular galaxies and along the arms of spirals. It includes the stars called blue giants, blue supergiants, and red supergiants. The regions where these stars occur are characterized by the presence of interstellar gas and dust and by bright regions of glowing hydrogen gas, ionized by the ultraviolet radiation of hot stars.

Type II is found in elliptical and lenticular galaxies and in the nuclear regions of spirals. It includes red giant stars, subgiant stars intermediate between the giants and the dwarfs, and probably also subdwarf stars.

Types I and II are usually mixed in galaxies. At first, astronomers believed that the ellipticals are composed mostly of type II stars and the magellanic irregulars of type I stars. However, later studies indicated that the diversity of stellar populations is probably greater than was first realized by astronomers.

DISTANCES OF GALAXIES

In order to determine the distances of galaxies outside of our own, we must first set up an absolute distance scale for stars in our own galaxy. The direct method of measuring distances by parallax would be of little use here, since it would apply only to stars within a radius of about 300 light-years—and our galaxy is about 80,000 light-years in diameter. Hence we have to adopt indirect methods, based on the proper motions, or apparent drift, of stars in the galaxy, their radial motions, their apparent magnitudes, and so on. By means of these methods, we can establish the distance and the intrinsic brightness of various

types of stars in our galaxy. If similar star types can be recognized and observed in other galaxies, they can be used as indicators of distance.

The extragalactic distance scale has been revised several times since 1952 and is still provisional. The following values for the nearer galaxies are often adopted:

| | <i>Distance in millions of light-years</i> |
|-------------------|--|
| Magellanic Clouds | 0.2 |
| Andromeda group | 2 |
| Ursa Major group | 8 |
| Virgo cluster | 39–52 |

For more distant galaxies, astronomers can make only rough estimates, ranging up to several hundred millions of light-years. The faintest galaxies that can be observed with the most powerful telescopes may be several thousand million light-years away.

DIMENSIONS OF GALAXIES

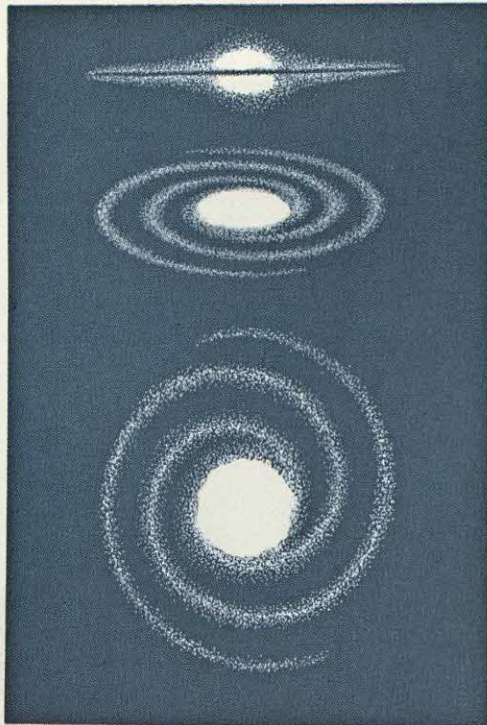
Once the distance of a galaxy is known, its intrinsic (actual) dimensions can be derived from its apparent dimensions, measured on a photographic plate. However, since the galaxies do not have sharply defined boundaries, it is difficult to determine these dimensions exactly. To compare the dimensions of galaxies of various types, it is necessary to study a fairly large number of them. They must all be observed under the same conditions.

Generally speaking, galaxies vary in size from dwarf systems having diameters of 10,000 light-years or thereabouts, to giant systems with diameters ranging up to 100,000 light-years. Dwarf galaxies are many times more numerous than the giants.

SPECTRAL TYPES AND COLORS

The first spectrogram of a galaxy, the Andromeda nebula, was obtained in 1899 by the German astronomer J. Scheiner at Potsdam Observatory. It showed absorption lines similar to those found in the spectrum of the sun. In addition to absorption lines, the spectra of some galaxies show bright emission lines, like those of gaseous nebulae in our own galaxy.

The system of spectral classification



The flatness of galaxies seen edgewise and their spiral arms suggest their rotation.

adopted for stars was extended to galaxies by W. W. Morgan at Yerkes Observatory. Of course the spectrum of a galaxy is a composite of the spectra of all the stars it contains. The following main spectral types can be recognized: B, A, F, G, and K, corresponding respectively to blue, white, yellow, orange, and red stars. These different types correspond to variations in the stellar populations of galaxies.

ROTATION AND MASSES OF GALAXIES

The flatness of many galaxies, when seen edgewise, and the presence of spiral arms suggested very early that the stars that composed them were rotating around the nucleus or center of these systems. The rotation of a galaxy can be observed and measured by the displacement of the lines in its spectrum, compared with reference lines in the spectrum of a fixed terrestrial object. If the galaxy is inclined to the line of sight at an angle of less than 90° , one side is moving away from the observer, and the lines of this section are displaced toward the red part of the spectrum. The other side is approaching, and the lines are displaced toward the blue part of the spectrum. This



Mt. Wilson and Palomar Observatory Photo



Harvard College Observatory

Two types of galaxies: a barred spiral galaxy (above), identified as NGC 1300, and the Large Magellanic Cloud (below), an irregular galaxy.

is called the Doppler effect, after its discoverer, the Austrian physicist Christian Doppler. The over-all effect of the rotation results in a shift of the spectral lines except in the center of the galaxy. For a galaxy seen face-on, the rotation cannot be detected by an examination of the lines in its spectrum.

The velocity of rotation of any part of a galaxy is affected by its distance from the nucleus. The period of rotation increases outward from a few million years near the nucleus to a few hundred million years in the outer region. Stars, interstellar gas, and dust all partake in the rotation, with about

the same velocity.

If we can determine which side of a spiral galaxy is nearer to the observer, we can determine the direction of rotation of the spiral arms. In all the cases studied, the arms were found to be trailing in the rotation.

We have just pointed out that the velocity of rotation is affected by the distance from the nucleus. If we analyze the variations in velocity, we have a method for determining the masses of galaxies. They range from a few thousand million times the mass of the sun, for dwarf systems, to several hundred thousand million times the solar mass, for giant systems.

PAIRS, GROUPS, AND CLUSTERS

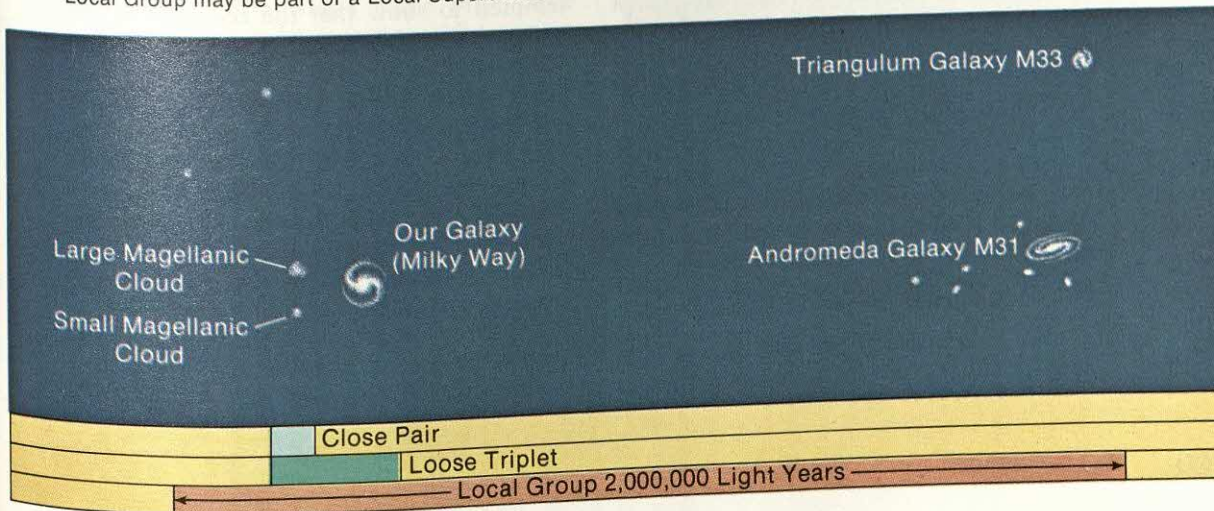
Galaxies occur frequently in pairs, triplets, and larger groups. The Large and Small Magellanic Clouds, in the southern skies, form a close pair, which is associated with our own galaxy to form a loose triplet. The Andromeda nebula (M 31) is the major member of a close triplet that includes M 32 and NGC 205. A score of nearby galaxies are members of the so-called Local Group of galaxies. Our own galaxy, M 31 (the Andromeda nebula), and the Magellanic Clouds form part of this group, which is about 2,000,000 light-years across. More

distant groups have also been observed. In many regions of the sky, galaxies form circular or elliptical clouds of roughly uniform density. In irregular, loose groups the great majority of the members in these clouds are spiral galaxies, with very few ellipticals or lenticulars. In dense groups, the elliptical and lenticular galaxies predominate.

In some areas, there are huge clusters, consisting of hundreds and often thousands of galaxies. Such is the Virgo cluster which, at a distance of about 40,000,000 light-years, has a diameter of about 7,000,000 light-years. At still greater distances, even larger and denser clusters have been observed. The great majority of members of these big clusters are elliptical and lenticular galaxies. The nearest of these, the Coma cluster, is at an estimated distance of 120,000,000 light-years. It has a total population of more than 10,000 galaxies in all and a general diameter of from 20,000,000 to 25,000,000 light-years.

Evidence has been obtained that at least certain clusters form still larger groups, called superclusters. Evidence has been presented for the existence of a Local Supercluster. This flattened system seems to include about a thousand of the brighter clusters of galaxies and probably several thousand of the fainter ones. The Virgo

Galaxies often occur in pairs, which may form part of larger groups. The larger groups may, in turn, be parts of still larger groups. For example, the Large and Small Magellanic Clouds form a close pair of galaxies. This pair is associated with our own Milky Way galaxy to form a loose triplet. The members of the triplet and various other galaxies, including the Andromeda Galaxy, make up the Local Group. The Local Group includes about 20 galaxies and is about 2,000,000 light years across. There is evidence that this Local Group may be part of a Local Supercluster.



Cluster is at the center of the system. The Local Group, of which our own galaxy is a member, is apparently not far from the edge. The overall diameter of the Local Supercluster is about 130,000,000 light-years; its thickness, about 30,000,000 light-years. The flattening of this supercluster suggested that it might be rotating. The rotation was confirmed in 1958 through an analysis of the velocities of several hundred bright galaxies.

INTERACTIONS BETWEEN GALAXIES

The close grouping of galaxies occasionally brings out spectacular interaction effects between neighboring systems. These effects have been specially investigated by F. Zwicky, at the Palomar Mountain Observatory. He found that they take a great variety of forms, depending on the distance between the galaxies, their sizes, their masses, and probably various physical properties still little understood. Very often, ribbonlike filaments of matter stream out from one galaxy to another. A filament may also emerge in the opposite direction. In certain cases, the outer arm of a spiral joins with a corresponding arm of a neighboring galaxy. When two galaxies are in collision and intermingle, vast antennalike streamers emerge from a chaotic central mass. In a few instances, galaxies which seem isolated display various distortions for which no visible companion can be considered responsible.

The mechanism of these interactions is not yet clearly understood. The presence of bright emission lines in the spectra of interacting galaxies indicates an unusual state of excitation of the interstellar gas. But no emission lines have been observed in the extended filaments and appendages of weakly interacting galaxies. This suggests that the luminosity of these filaments is due to starlight and not to the excitation of interstellar gas. It is difficult to understand how such long filaments of matter can remain stable for any length of time.

RED SHIFTS AND RECEDING GALAXIES

We have already pointed out that the Doppler effect provides a means for deter-

mining whether a star or group of stars is receding from an observer or whether it is approaching him. The first radial—that is, line-of-sight—velocities of galaxies were measured between 1917 and 1926 by the American astronomer Vesto M. Slipher at the Lowell Observatory in Arizona. Except for a few of the nearest galaxies, they were recession velocities; the galaxies were moving away from the earth. In 1929, Hubble discovered that the greater the distance of a galaxy, the greater its recession velocity.

Galaxies are speeding away from us at a truly awesome rate. Milton L. Humason, at Mount Wilson Observatory, measured velocities of up to 40,000 kilometers per second. Still greater ones have been observed with the 500-centimeter telescope on Palomar Mountain. They range up to 61,000 kilometers per second, or more than one-fifth the velocity of light, for members of a distant cluster of galaxies in the constellation Hydra. The American astronomer William A. Baum has estimated that certain velocities are in excess of 150,000 kilometers per second, or about one-half the speed of light. In all the cases we have mentioned above, it has been found that the recession velocity of a given galaxy is very nearly proportional to distance.

THEORY OF AN EXPANDING UNIVERSE

The belief that galaxies are receding is used as evidence for the theory that the universe is expanding. This theory holds that the universe began with a "big bang" explosion and that it has been expanding ever since. Some astronomers have attempted to show that the red shift of the spectral lines of the galaxies is not due to their movement away but rather to some other effect. These astronomers, who are really questioning the validity of the Doppler effect, have not, however, been able to provide proof for a different theory. Therefore, the theory of the expanding universe is still widely accepted. We may think of space itself as expanding and in the process carrying the galaxies away with it. As viewed from any galaxy, the effect would be the same. All other galaxies would appear to be receding.

COSMIC RAYS

by Volney C. Wilson

Whizzing through our galaxy at all times, in every direction and at fantastic speed, are the charged particles that we call cosmic rays. They are really atomic nuclei—that is, atoms stripped of their electrons. Traveling at nearly the velocity of light, some have energies far greater than those produced with powerful atom smashers.

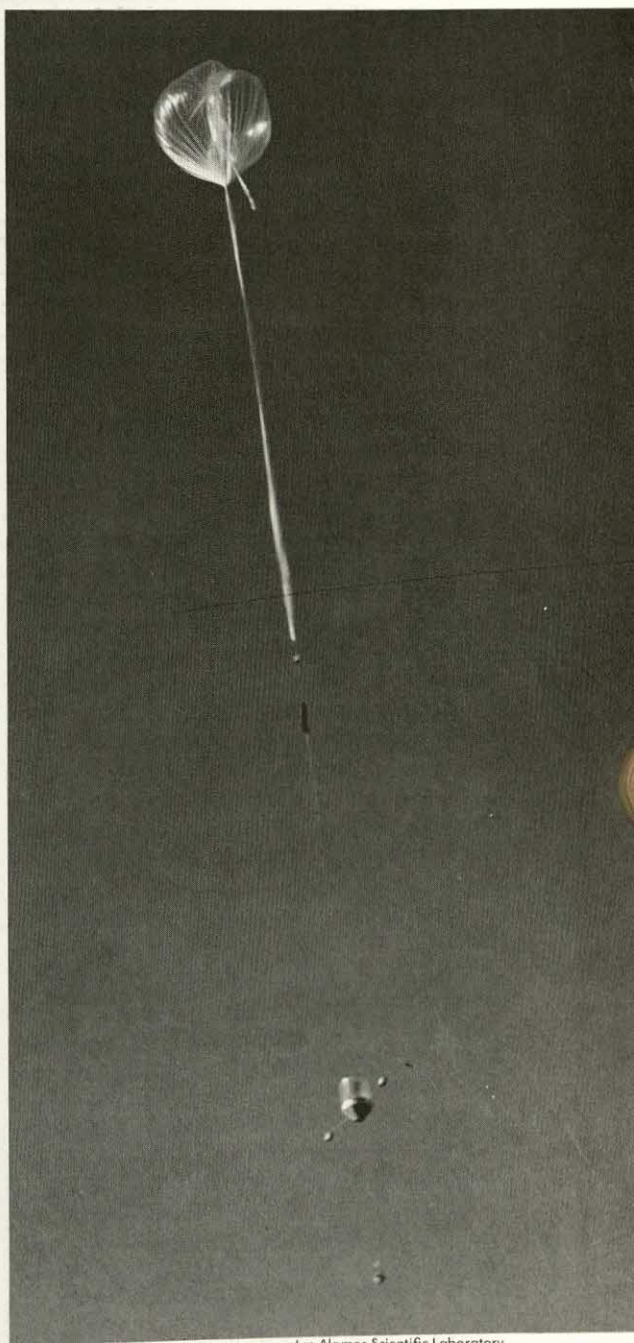
The cosmic rays that fly through outer space are called primary cosmic rays. A thin rain of these particles constantly strikes the upper atmosphere of the earth. As they enter the upper part of the atmosphere, they collide with atoms of air. The fragments known as secondary cosmic rays result from the collisions.

About one per cent of these secondary rays penetrate the remaining atmosphere and reach sea level. Roughly 600 rays pass through the human body every minute, day and night. This should not alarm us, though. A wristwatch with a radium-painted face exposes us to the same dose of radiation. Doctors estimate that the dose must be 30 to 300 times greater to produce harmful genetic effects on human beings by increasing the rate of mutations.

If we could harness the energy of all cosmic particles striking sea level over the entire world, it would yield little more power than a modern automobile engine. Cosmic rays will never power our machines; but they do provide nuclear physicists with a natural laboratory for the study of super-high-energy phenomena. They may reveal the secrets of the forces that bind together the particles of matter in atomic nuclei. They may tell us about the universe beyond our solar system.

DETECTION OF COSMIC RAYS

Cosmic rays made their presence felt long before they were identified. About the



Los Alamos Scientific Laboratory
University of California

In Operation Skyhook, a balloon carried cosmic-ray detecting equipment to high altitudes. Cosmic rays strike photographic and plastic emulsions in the detector and leave a record of their visit. An analysis of these records reveals the nature and intensity of the rays.

beginning of the twentieth century, physicists discovered that electroscopes and ionization chambers were affected by certain mysterious rays existing in the atmosphere. An electroscope is a device for determining the electric charge on a body. An ionization chamber is used to measure the intensity of radiations. These rays were far more penetrating than X rays or any other forms of radiation known at the time. High-energy X rays could be stopped by a lead plate only 1.6 millimeters thick; but 10 centimeters of lead would absorb only 80 per cent of the "new" rays.

The first and most natural assumption was that these penetrating rays issued from some special radioactive material in the earth's crust and atmosphere. Beginning in 1909, however, enterprising experimenters carried ionization chambers aloft in balloons and found the rays grew more intense with increasing altitude. Between 1910 and 1914, Victor F. Hess of Austria and Werner Kolhörster of Germany made the first precise measurements, at altitudes up to 9,000 meters. At this height the rays were ten times more intense than at sea level. Clearly, the penetrating rays traveled downward through the atmosphere. Hess came to the conclusion that the mysterious rays must originate in outer space.

In 1925, a series of experiments by a team of physicists headed by Robert A. Millikan, of the California Institute of Technology, confirmed Hess's hypothesis. Millikan and his colleagues lowered ionization chambers into two snow-fed lakes at high altitudes in California in order to determine the absorption of the penetrating rays in water, as compared with their absorption in air. Measuring the intensity of the rays under different depths of water and at different altitudes, they came to the conclusion that the rays must originate outside the atmosphere. In the report of these experiments, the name "cosmic rays" was used for the first time.

In a few years, two newly invented instruments revealed important new facts about cosmic rays and later became the cosmic-ray physicist's chief tools. In 1927, the Russian physicist D. V. Skobeltsyn first

adapted the Wilson cloud chamber to the study of the rays. In the Wilson cloud chamber, the path of an ionizing particle or ray is made visible as a trail of water droplets.

In 1928 and 1929, Werner Kolhörster and another German, Walther Bothe, devised a research technique using sets of Geiger-Mueller counters, devices that detect particles by electrical discharges. By arranging the counters in a straight line and providing them with the proper electrical circuits, one could trace the path of a single cosmic ray.

The bubble chamber, developed in 1952 by Donald Glaser, also proved helpful in the analysis of the rays. In the bubble chamber, a liquid is kept just below the boiling point. The pressure is suddenly lowered and the liquid becomes superheated. When high-speed charged particles now pass through, their passage will be indicated by a series of bubbles.

By the use of these devices, it has been established that most of the secondary cosmic rays are high-energy, electrically charged particles.

COSMIC RAYS AND EARTH'S MAGNETIC FIELD

In 1927, the Dutch scientist Jacob Clay discovered that primary cosmic rays are affected by the earth's magnetic field. During a voyage between Amsterdam and Indonesia, he observed that the intensity of cosmic rays drops as one approaches the magnetic equator from higher latitudes. The existence of this "latitude effect" seemed to show that primary cosmic rays are electrically charged particles.

A neutral (uncharged) particle moving in a straight line through a magnetic field is not influenced by the field. However, a charged particle in the same situation will have its path bent into a curve, if it moves approximately at right angles to the magnetic lines of force.

The earth's magnetic field is very weak, but it extends for thousands of kilometers into space. Any charged particle approaching the earth must travel immense distances through the field and will be ap-

preciably deflected by the curving force. This force is greatest on particles that approach exactly at right angles to the lines of force, and weakest on those that travel parallel with the lines. The bending is more pronounced for slow-moving, or low momentum, particles than for faster, or high momentum, ones.

This is illustrated in the diagram on page 226. The dotted lines represent the earth's magnetic lines of force. *A*, *B*, *C*, *D* and *E* are all charged particles approaching the earth with the same momentum or energy. *A* is only slightly deflected by the field because it approaches near the pole and thus moves almost parallel with the lines of force. On the other hand, *E* approaches near the equator, moving exactly at right angles to the lines. The strong curving force on this particle eventually turns it back into

the direction from which it came. The paths taken by *B*, *C* and *D* will depend upon the angle formed by their direction as they approach the earth and the magnetic lines of force. For a particle to penetrate the magnetic field and the atmosphere at the equator (particle *F*), it would have to start out with many times more momentum than the other particles.

COSMIC-RAY INTENSITY

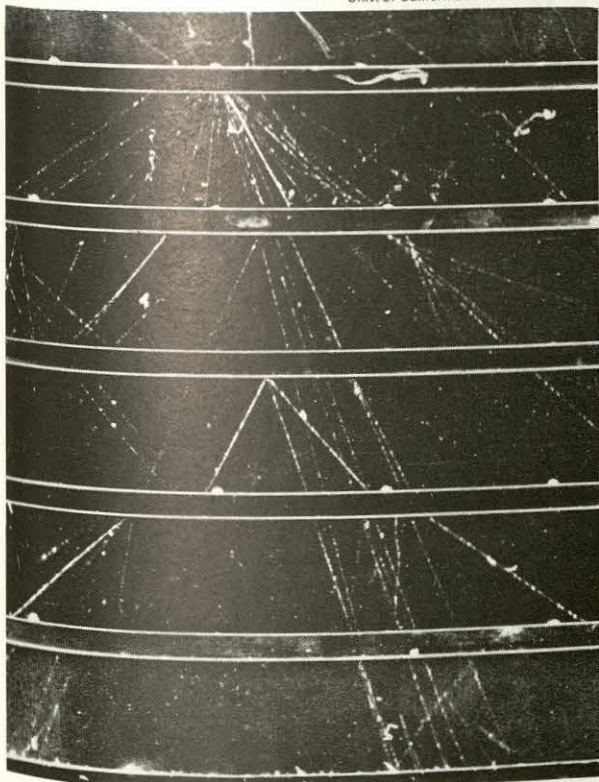
If the particles are indeed electrically charged, we should expect many fewer rays to reach the earth's surface at the equator than at higher altitudes. This was conclusively shown. In 1930, a worldwide survey was begun, under the direction of Arthur H. Compton, of the University of Chicago, to find out cosmic-ray intensities at different latitudes and altitudes all over the globe. The report of the survey in 1933 established that, from the geomagnetic latitudes of 50° north (or south) to the equator, cosmic-ray intensity at sea level drops about ten per cent. It showed that primary cosmic particles are electrically charged.

Is the charge positive or negative? Manuel S. Vallarta, a Mexican mathematician, calculated in 1933 that a positive particle with low momentum could reach the earth more easily when approaching from the west than from the east. The directions would be reversed for a negative particle. By 1938, experiments proved that the rays falling from the west were definitely more intense than those from the east. This led to the conclusion that the primary cosmic rays are positively charged.

The American physicist Scott E. Forbush discovered that cosmic-ray intensities decrease during periods of high sunspot activity. The sun is continuously throwing out tremendous quantities of protons moving at a speed of about 1,600 kilometers per second—a phenomenon referred to as the solar wind. During periods of sunspot activity, when a solar flare happens to be at a particular place on the sun's surface, the solar wind in the direction of the earth is greatly increased. This in turn increases the strength of the magnetic field in the vicinity of the earth to such an extent as

Cosmic-ray particles have penetrated a cloud chamber with a series of lead plates across it. As the particles pass through the lead plates, they are slowed down and split, producing a cosmic-ray shower.

Univ. of California Radiation Laboratory



to shield out some of the low-energy cosmic rays. This is called the Forbush effect.

Rockets and artificial satellites have been used increasingly to measure cosmic-ray intensity above the earth. They have also served to analyze the effect of the solar wind upon the earth's magnetic field.

COSMIC-RAY SHOWERS

The Italian-born physicist Bruno Rossi showed in 1932 that if three or more Geiger-Mueller counters were spread out horizontally in an irregular pattern, they would sometimes be tripped simultaneously, indicating that showers of particles were traveling together. These showers could be produced as cosmic-ray particles passed through lead or other substances containing heavy atoms. Rossi concluded that each shower was produced from a single cosmic-ray particle as it passed close to the nucleus of a nearby atom. Showers could also be observed in a Wilson cloud chamber that had a lead plate across it. The track of a single cosmic-ray particle is seen entering the chamber. As it passes through the lead plate, the track splits in two. These tracks split again and again, producing a shower of particles.

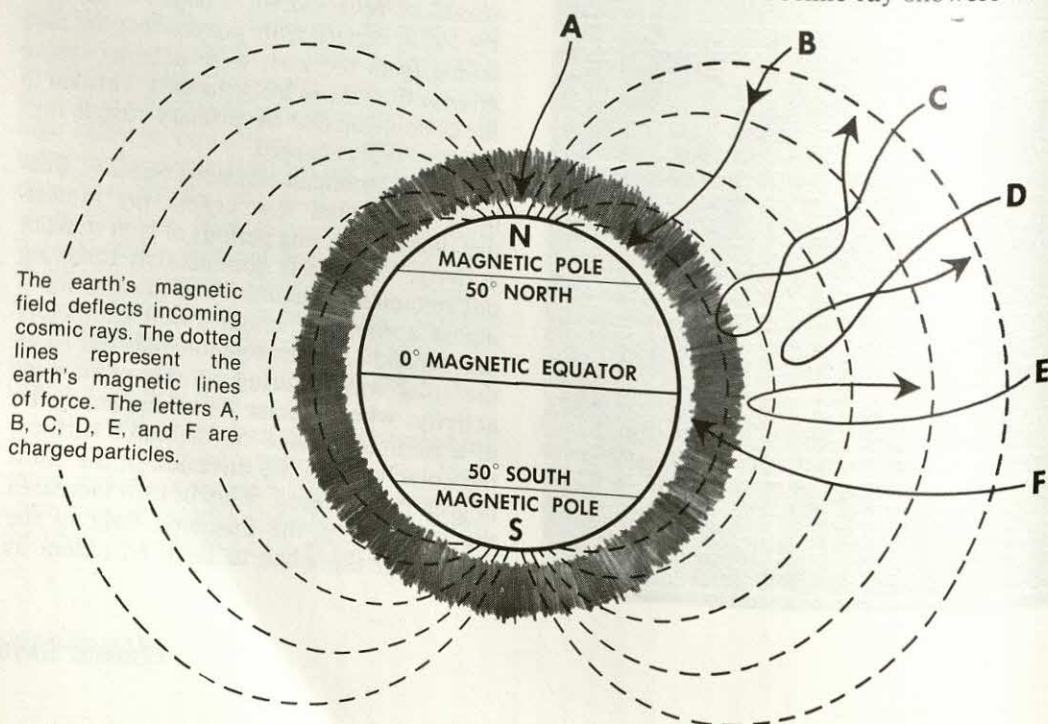
In the same year in which Rossi revealed the existence of cosmic-ray showers, the American Carl D. Anderson dis-

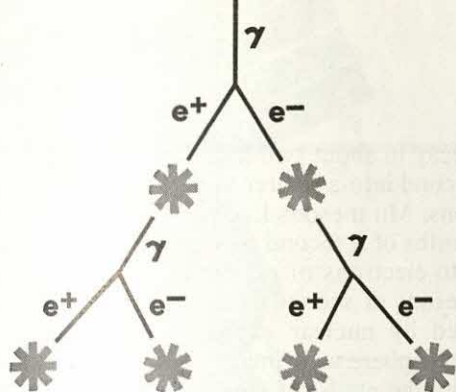
covered a new fundamental particle in cosmic radiation—the positive electron, or positron. It appeared in a photograph of a cloud chamber containing a lead plate in a strong magnetic field. Anderson's historic achievement won him a share of the 1936 Nobel Prize in physics.

Cloud-chamber photographs soon showed that each time a shower track split, the two branches were oppositely curved. An electron and a positron were created. A pair-production theory of cosmic-ray showers was now formulated. It was maintained that the process of shower formation represented the conversion of energy into charged matter and vice versa.

This is what takes place according to the pair-production theory. As a gamma-ray photon—a fragment of light energy—penetrates the nuclear field of an atom, a portion of its energy is suddenly transformed into a pair of electrons, positive and negative. The remaining energy provides the velocities of the electron and the positron. Each of the daughter electrons may disappear in its turn, producing another gamma ray, if it is suddenly slowed down in the field of another nucleus. Two gamma rays may produce four electrons; four gamma rays may then arise, then eight electrons and so on.

In 1938 such cosmic-ray showers were





The pair-production theory of cosmic-ray showers. A gamma ray (indicated by γ) yields a pair of electrons, positive and negative, indicated by e^+ and e^- , respectively. Each of these electrons may disappear, producing another gamma ray as it does. Each of the gamma rays in turn may yield two electrons. As the process is continued, a cosmic-ray shower is produced.

discovered in the atmosphere, up to one-third kilometer in diameter and with millions of particles each. The energy of one cosmic-ray particle causing a shower may reach 1,000,000,000 GeV. 1 GeV = 1 giga-electron volt, or 1,000,000,000 electron volts. An electron volt is the energy gained by an electron when accelerated by one volt of electricity. In contrast, fission of a uranium atom releases only 200 MeV. 1 MeV = 1,000,000 electron volts.

THE PRIMARY COSMIC RAYS

Primary cosmic rays contain 91 per cent positive particles—protons, which are the nuclei of hydrogen atoms; 8 per cent alpha particles, or helium nuclei, with 2 protons apiece; and one per cent nuclei of heavier elements. The latter are mostly lithium (3 protons per nucleus), beryllium (4 protons), boron (5 protons), oxygen (8 protons), and iron (26 protons). The particles are detected by photographic emulsions in high-altitude balloons.

Heavier cosmic-ray nuclei are being found. Scientists have learned that moving heavy particles damage mica and plastics. They leave tracks, which, when etched, give an idea of the particles' sizes and charges. One nucleus has been found to contain 106 protons, representing an element more massively charged than uranium, which has 92 protons per nucleus.

Cosmic-ray composition is similar to the known distribution of chemical ele-

ments in stars, nebulae, and interstellar dust. Cosmic-ray composition and the distribution of the quantities of elements in the universe show sharp differences in the numbers of atoms having odd and even numbers of protons per nucleus, with a peak in regard to iron. "Even-numbered" nuclei are more common. These facts give a clue to the origin of primary cosmic rays.

TWO COMPONENTS OF SECONDARY COSMIC RAYS

Secondary cosmic rays are composed of two very different classes of particles, which can be separated by a piece of lead about 13 centimeters thick. One component is completely absorbed in this thickness. This soft component consists mainly of electrons, positrons, and gamma-ray photons—the typical particles of a cosmic-ray shower. The intensity of this soft component increases from the top of the atmosphere down to an altitude of about 17 kilometers, where it makes up roughly four fifths of the total radiation. From this altitude downward, the intensity decreases until it makes up only one quarter of the total at sea level.

The other component passes through 13 centimeters of lead almost unobstructed. It is called the hard, or penetrating, component. It decreases in intensity continuously from the top of the atmosphere down to sea level. Approximately one half of these rays at sea level can still penetrate 38 centimeters of lead. The difference in the absorption pattern and penetrating power of the two components represented a puzzling problem to physicists for many years.

In 1936, the riddle was finally solved by the teams of C. D. Anderson and S. H. Neddermeyer, at the California Institute of Technology, and J. C. Street and E. C. Stevenson, at Harvard University. Working independently, these teams showed that all the evidence pointed to a new type of fundamental particle, intermediate in mass between the proton and the electron. This was the mesotron, now called the meson.

In the meantime, the author of this article had been measuring the penetration of cosmic rays below the ground in a copper

mine in northern Michigan. He reported that some rays were able to penetrate as much as 500 meters of rock. Further investigation showed that these rays were corpuscular, or made up of tiny particles, and that they ionized the rock and lost energy all the way down. Calculations showed that they must be the newly discovered mesons. It is now agreed that mesons make up the bulk of the hard component of secondary cosmic rays. The high energy of mesons accounts for the penetrating power of this component.

The existence of the meson had been predicted by the Japanese nuclear physicist Hideki Yukawa in 1935. He also predicted that the meson would be unstable outside the nucleus, decaying with the emission of an electron. In 1940, this decay was actually photographed for the first time in a Wilson cloud chamber.

In 1948, researchers at Berkeley, California, first succeeded in producing mesons artificially, by bombarding carbon atoms with alpha particles accelerated to 380 MeV. Both positive and negative mesons were produced. Some, known as pi mesons,

Underground cosmic ray detectors have been built to test theories of the origin of cosmic rays. Here a physicist connects wires from 300 cosmic ray counters to a central computer housed in a nearby mine.

University of Utah



decay in about two hundred millionths of a second into a lighter variety called mu mesons. Mu mesons live for roughly two millionths of a second before they disintegrate into electrons or positrons. The heavier pi mesons of secondary cosmic rays are created by nuclear explosions in the upper atmosphere as primary cosmic rays collide with the nuclei of atoms—the atoms of the gases that make up the atmosphere. The pi mesons usually decay in flight and produce mu mesons, which live long enough to travel great distances. It is the mu mesons that form the hard component of secondary cosmic rays at sea level.

ORIGIN OF COSMIC RAYS

What is the source of cosmic rays and how do they acquire their fantastic energies? Some astronomers have long held that the universe was created in a single primordial explosion, and that cosmic rays are tiny remnants of this colossal event. But the presence of heavy nuclei in the primary rays shows that their energies must be acquired gradually. In an explosive process, the larger nuclei would be completely shattered.

Present evidence suggests that most cosmic rays originate and remain largely within our own galaxy. Let us suppose that the distribution of rays observed in our galaxy extends far beyond its borders throughout the universe. We know the average energy of the rays in our region. From this we can calculate the total energy carried by cosmic rays everywhere. This calculation has been performed. The figure reached would equal the entire mass of the universe if this were converted into energy. Scientists consider such an idea impossible. Moreover, the solar system is constantly rotating along with the rest of our galaxy. If rays are entering the galaxy from outside, certain complex effects due to the rotation of the galaxy should be visible. None have ever been observed.

At least 90 per cent of the cosmic rays originate inside our own galaxy. The maximum lifetime of a cosmic ray is about 2,000,000 years. The known characteristics of cosmic rays suggest they originated



in thermonuclear (atomic) explosions in certain kinds of stars.

Old stars rich in iron atoms reach a stage when they just blow up. They then release vast amounts of particles and electromagnetic radiation, including light. These explosions occur by a process called neutron capture. Neutrons—heavy nuclear particles with no charge—can move freely outside the nuclei of atoms. They may, however, be absorbed, or captured, by the nuclei of atoms; this event makes the atoms unstable. The atomic nuclei, therefore, release energy in the form of alpha particles, protons, neutrons, and radiation of very short wavelengths, especially gamma rays, in a gigantic blast.

An exploding star is called a supernova. It is much brighter than an ordinary nova, which does not really explode, but shoots out matter instead. The best-known supernova was the event of 1054 A.D., recorded by Chinese astronomers. The remnant of this nova is the Crab Nebula.

At present, the Crab Nebula is observed to emit radio waves and light that is markedly polarized. That is, the radiation waves vibrate along a single plane or along a few planes, instead of at many angles, as ordinary (unpolarized) radiation does. The polarization means that the Crab radiation originates from electrons moving at great speeds in strong magnetic fields. This is called synchrotron radiation. A synchro-

tronic cosmic-ray detector of the Orbiting Geophysical Observatory, an artificial satellite. Electrons in cosmic rays are measured by the rate of energy loss they undergo as they pass through this device. The electrons cause flashes in certain crystals. The flashes are changed into electrical pulses, which are analyzed and transmitted to earth.

NASA

tron is an atom-smasher, or accelerator, that speeds up electrons and other particles by means of changing, intense magnetic fields. Polarized electromagnetic radiation is observed to arise from electrons accelerated in a synchrotron.

The Crab Nebula, then, may act as a giant cosmic accelerator of particles—or cyclotron—with energies that dwarf those of man-made accelerators. The most powerful man-made accelerator to date reaches only 500 GeV. Some cosmic-ray energies reach the staggering figure of 1,000,000,000 GeV!

Within the Crab Nebula astronomers have identified a pulsar—a very small and very dense star, composed mostly of neutrons, that emits radiation in very short bursts, or pulses.

The pulsar itself is part of the star that was seen to explode over 900 years ago. This pulsar may be the chief energy source of the Crab Nebula. Not only does it radiate energy, but it ejects large numbers of particles from its surface. These particles are being accelerated in the very strong rotating magnetic field of the pulsar. Scientists estimate that one to two supernovae occur in our galaxy per century. If they contain pulsars, the energy balances are such that these supernovae could just about produce the observed cosmic rays. Rays with energies of 1,000,000,000 GeV or less are trapped in the magnetic field of our galaxy (the Milky Way). They last an average of about 1,000,000 years before striking an atom and being annihilated in the process.

The extremely few cosmic rays with energies greater than 1,000,000,000 GeV may have arrived at the earth before escaping the galaxy or may even have come from outside the Milky Way. Observations so far suggest that no more than ten per cent, and perhaps far less, of cosmic rays come from beyond our galaxy.

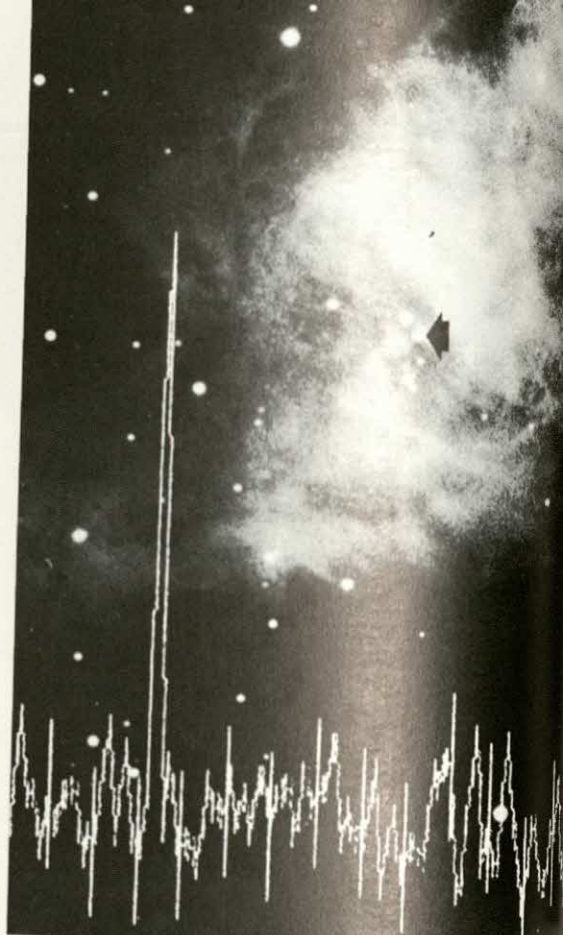
PULSARS

by Antony Hewish

The discovery of pulsars in 1967 was one of those lucky accidents that sometimes happen in scientific research. I only wish I could say that we were looking for pulsars at the time. But the truth is that my colleagues and I were studying quasars—mysterious radio objects currently thought to be galaxies, situated far beyond the confines of the Milky Way—when the first pulsar unexpectedly placed its signature upon our records. By an extremely fortunate twist of fate, the new radio telescope that we were using was ideally suited to pick up the rapid successions of faint radio pulses that characterize these fascinating objects.

It all began in 1964 when we noticed that certain radio galaxies, noted for their tiny angular extent, exhibited rapid and irregular variations of intensity. This phenomenon was very similar to the "twinkling" of visible stars, and we eventually learned that it was caused by clouds of hot gas ejected from the sun. These clouds take part in the violent and continual outflow of material from the sun known as the "solar wind." Because the clouds are so hot, electrons are stripped from the atoms, and the gas becomes a mixture of electrically charged particles (electrons and positive ions). Radio waves are deflected slightly as they pass through these clouds; as a result, originally parallel rays from a distant radio galaxy become entangled as they traverse the solar system. It is this effect that causes the radio galaxies to "twinkle." Only very compact sources like quasars are subject to such "twinkling"; ordinary radio galaxies tend to average out the fluctuations.

We thought that this "twinkling effect" would be an excellent method of finding, among the several thousand radio galaxies that have been charted, those galaxies that actually *are* quasars. So in 1966 we started to build a radio telescope especially designed for this purpose. It was located next to Sir Martin Ryle's famous "one mile" ra-

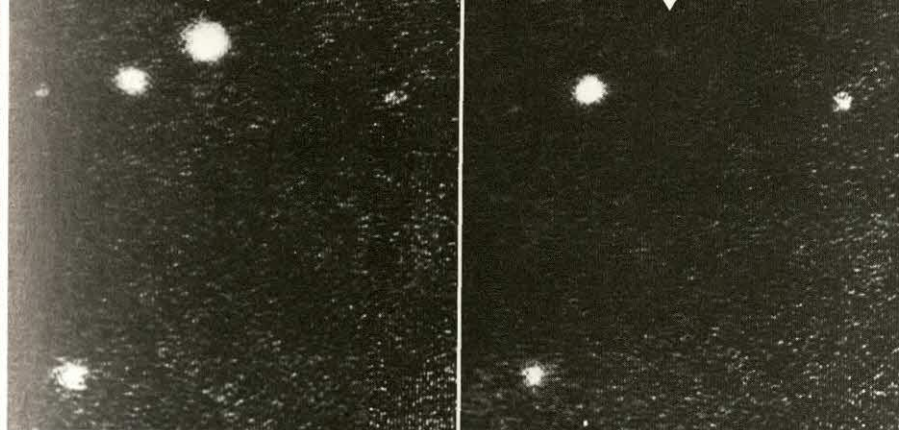


Mt. Wilson Observatory, Naval Research Lab., D.C.; "Science"

dio telescope outside Cambridge, England, and consisted of a huge array of dipoles—more than 2,000 in all. (A dipole is an antenna consisting of two rods extending in opposite directions.) The array was very sensitive, since it collected the incoming radio waves over an area of almost 2 hectares and operated at a wavelength of 3.7 meters.

A TWINKLING AT NIGHT

Observations with the new radio telescope began in July 1967. It took one week to scan a large fraction of the sky, and exactly similar observations were repeated, week by week, to investigate how the radio "twinkling" of the quasars varied in strength as the sources were studied at different angular distances from the sun. The results were used to give us a quantitative measure of the angular sizes of the quasars. The radio signals were recorded as an ink trace on a moving paper chart. Approximately 75 meters of chart had to be inspected each



Lick Observatory

Above left: the arrow points to the pulsar in the Crab Nebula. It was once thought to be an ordinary star like the object to its left. Above right: the same area of the sky, photographed a fraction of a second later. The pulsar has "turned off."

A photograph of the Crab Nebula. The arrow points to the optical pulsar NP 0532. The power spectrum shows the distribution of X ray pulses from the nebula. The peaks represent the "beat" of the pulsar. These X ray pulses occur at the same rate as the star's optical and radio pulses—33 times a second.

week. This task was carried out by Miss Jocelyn Bell (now Dr. Burnell), a young graduate student from Ireland who had also worked on the construction of the array.

One day in August, Miss Bell showed me an odd-looking tracing on which a radio source had apparently been "twinkling" in the middle of the night. This was unusual because intensity variations caused by the solar gas are always very small when observations are carried out in directions away from the sun. We did not really believe that the signals were genuine, because radio telescopes often pick up terrestrial interference of one kind or another (such as that from automobile ignitions). We made a note, however, of the apparent direction in the sky from which these radio signals had originated.

In the weeks that followed, the strange signals sometimes reappeared, but often were not recorded at all. By September, however, it was plain that the phenomenon could no longer be ignored. The next step

was to obtain a detailed recording, using a much faster chart speed in order to make a closer inspection of the signals. Yet no sooner was the equipment set up than the source faded out. Not until late November did it return, and when Miss Bell showed me the first recording, I frankly could not believe what I saw. The signals from this mysterious source were in the form of short-duration pulses spaced at intervals of about $1\frac{1}{3}$ seconds. Of course we had to check this on several successive days to convince ourselves that we were not being tricked in some way. But the same signals were always there, and during the first week of December, the source flared up in intensity and produced some really beautiful pulses.

ARTIFICIAL LOOKING SIGNALS

The period that followed was one of intense excitement. What astronomical body could be the origin of such artificial-looking signals? From the lack of any detectable parallax, or the small change in the apparent position of a star when measured at different times of the year, we knew that the object was far beyond the solar system. We also knew that the body had to be small—probably about the size of the earth; a large body simply cannot emit a short-duration pulse, because radiation leaving different parts of its surface will arrive at different times. The possibility that we were, for the first time, in contact with intelligent beings elsewhere in the galaxy could not be ignored.

Naturally we had to press these investigations a little further before releasing the news of our discovery. Our first task was the accurate timing of the pulses. Within a few days we found that the pulses were keeping time to better than one part in a million; whatever the nature of the object, it was acting like a very-high-quality clock. After a month it was clear that the signals could not be originating from a planet. If they had been, we would have detected a systematic variation in the period of the pulses owing to the assumed planet's orbital motion. We also made a careful search of the existing survey recordings for similar pulsing sources, and further observations confirmed the presence of three more in widely differing parts of the sky. At this stage we felt confident that we were dealing with a natural phenomenon, probably some heretofore unknown kind of star.

We announced our findings to the scientific community.

MILLION TON LIGHTHOUSE

The news of our discovery triggered an intensive period of activity, which was probably unique in the history of astronomy. Radio telescopes all over the world joined the quest for further observational data, and theoretical astrophysicists considered every conceivable possibility of accounting for the pulsars. One year after our first results were published, a total of 31 pulsars had been found. While it is still not certain what they are, the general belief is that they must be the long-sought neutron stars — predicted theoretically but not found until now. These tiny stars, perhaps only a few kilometers or so in diameter yet containing as much matter as our entire sun, are likely to be spinning at high speeds. One idea is that they behave like celestial "lighthouses," throwing out a beam that flashes round the sky at the rotation period of the star.

To understand why this seemingly far-fetched theory is popular, it is necessary to digress for a moment and consider the final stages in the life cycle of a star. What happens, for example, when a star has burned all its available fuel? Does it simply cool down and turn into a dark ball of solid ashes?

Definitely not! When we talk about a star "burning," we are referring to the nuclear-fusion process (as in a hydrogen bomb) in which hydrogen is converted to helium and other more complex nuclei. Vast quantities of intense radiation are released in these reactions, and the resulting radiation pressure keeps a typical star like the sun extended against the inward gravitational pull. After a certain time, however, the nuclear fuel is exhausted, and gravity causes the star to shrink. Various possibilities may now occur.

For a star like the sun, the initial shrinking under gravity produces energy that may puff the star out to become a red giant; but in the end it will collapse to a white dwarf. At this stage, when the star is no larger than a small planet, it is the compression of the electrons that eventually resists further gravitational collapse. The matter is so condensed that normal atoms do not exist; the negatively charged electrons are pressed close to the positively charged nuclei. A teaspoonful of the mixture would weigh several metric tons.

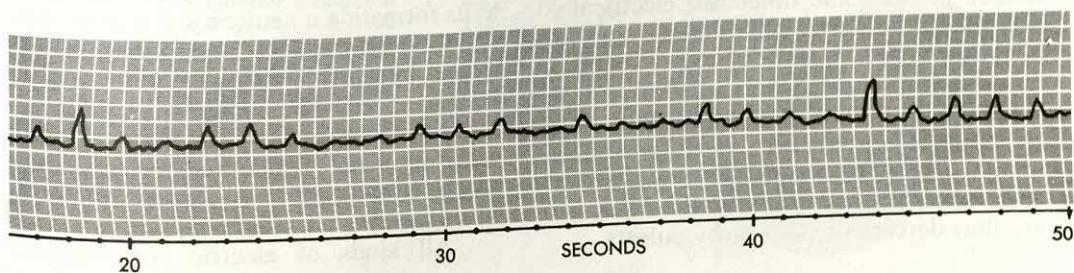
For a star that is heavier than about $1\frac{1}{2}$ times the solar mass, a rather more violent end is predicted. Initially, the central part of the star begins to collapse, as in a white dwarf. The inward gravitational force is so enormous that the electrons are, so to speak, squashed right inside the nuclei. They combine with protons to form neutrons, thereby ceasing to exist as independent particles. The neutrons, being electrically uncharged, can be packed much more closely than electrons. Calculations show that the matter density must increase to about one million metric tons per teaspoonful before the inward gravitational forces are balanced. By this time the inward "falling" matter is traveling so fast that when it is finally halted, a blast of energy rips outward, initiating a mighty explosion and blowing the outer shell of the star far out into space. This, astronomers believe, is the origin of a supernova. In the center of the resulting debris we expect to find a neutron star, a ball of fantastically dense matter, almost as heavy as the sun but only a few kilometers in diameter.

CRAB NEBULA

One of the main reasons why pulsars are thought to be neutron stars is a famous supernova witnessed by Chinese astronomers in A.D. 1054. Its remnants—the so-called Crab Nebula—contain a pulsar near the center, as predicted by theory. The Crab pulsar itself was found only in 1968, by the National Radio Astronomy Observatory, at Green Bank, West Virginia. Its period, measured with the giant reflector at Arecibo, Puerto Rico, is 0.033 second—making it the most rapidly pulsating radio source known.

This pulsar created excitement in 1969, when Steward Observatory in Tucson, Arizona, reported a visible (light-emitting) star in the center of the Crab Nebula. It was flashing at exactly the same rate as that at which the Crab pulsar was emitting radio signals. This visible star has been known for many years as the probable source of the supernova explosion. It is now identified with the Crab radio pulsar. Astronomers discovered that its light comes in pulses only because they used a light detector blinking at the same rate as the radio pulsar. X-ray pulses at the same rate have since also been received from the pulsar.

Right: a radio source that twinkles at night. These radio emissions are from CP 1919, the first pulsar to be identified. This recording and the one below were made by the author and his coworkers at the Mullard Radio Astronomy Observatory near Cambridge, England. The recording below shows the regular radio emissions of pulsar CP 1919. Fifteen pulses can be distinguished in each 20-second interval. This corresponds to one pulse approximately every 1.33 seconds. Some pulsars have much shorter periods; others emit signals at longer intervals.

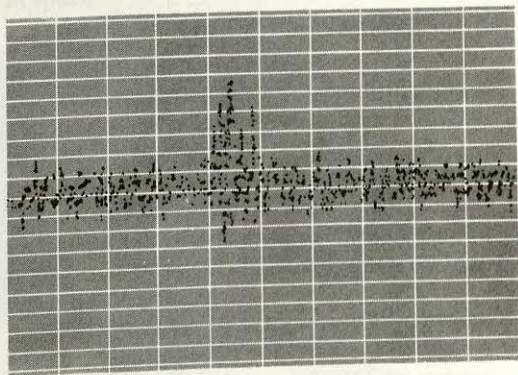


Do other supernovas contain pulsars? Estimates of the total number of pulsars in our Milky Way galaxy agree well with the number of supernova explosions. However, despite intensive search, only some of the fewer than a hundred known pulsars have been found to lie close to gas clouds that mark the sites of past stellar explosions. But the situation may be explained by the fact that most pulsars are probably at least several million years old. This is long enough for a pulsar to move far away from the place where it originated from a supernova.

AGE OF PULSARS

The ages of pulsars are estimated from accurate timing measurements, which show that, in general, their pulse rates are steadily slowing down. According to the aforementioned "lighthouse" theory, the decrease is due to the neutron star's spinning more slowly and losing energy. Thus, the age of the Crab pulsar is calculated at 1,000 years, which agrees well with the known calendar age of the Crab supernova (900-odd years).

The Crab pulsar is about 1,000 times younger than most pulsars. This fact ex-





SUN

EARTH

NEUTRON
STAR

PULSAR



plains why it pulses so fast and also why it is the only known pulsar that emits light and X rays as well as radio waves.

The discovery of the Crab pulsar has solved a longtime mystery about the Crab Nebula itself. The nebula is "lit up" by something that supplies continuous energy and that astronomers could not identify until very recently. This "something," of course, is the Crab pulsar. The energy of rotation lost by the pulsar as it slows is enough to energize the Crab Nebula. The agreement between the rate of pulsar's energy loss and the nebula's energy gain is a striking confirmation of the "lighthouse" theory.

DISTANCE OF PULSARS

The radio signals from pulsars also tell us something about their distances from the earth. We already know that the Crab pulsar is 6,000 light-years from us, by the study of its light. But the distances of invisible pulsars can only be estimated roughly from the different wavelengths of the pulses, which arrive on earth at slightly different times.

Pulsars emit radio signals at wavelengths of 0.1 meter to 7 meters. Pulses of varied wavelengths start from a given source at the same time. But electrically charged gas in space makes them travel at slightly different velocities.

The measured differences in arrival times of the pulses give the distance to a pulsar. Most pulsars turn out to be 100 to 10,000 light-years away. Since the diameter of our galaxy is over 60,000 light-years, we are thus detecting only nearby pulsars.

Relative sizes of the sun, earth, white dwarfs, neutron stars and pulsars. Pulsars are believed to be somewhere between neutron stars and white dwarfs in size. They are apparently much, much smaller than our sun, which has a diameter of about 1,390,000 kilometers (the earth's diameter is about 13,000 kilometers). However, dwarf stars, neutron stars and perhaps pulsars are very dense and weigh as much as the much larger sun.

THE ENERGY OF PULSARS

Knowing the distances to the pulsars enables us to calculate how much energy they are radiating. Except for the Crab pulsar, they are weak emitters compared with a common star like the sun. As yet, however, there is no accepted theory that accounts for the radio emission. One important feature of neutron stars is that they must have enormously powerful magnetic fields at their surface. This follows from the fact that ordinary stars are magnetized, just like the earth, and when a star collapses to the neutron-star condition, the associated compression of the magnetic field causes its strength to be magnified up to 10,000,000,000 times the original value. At its formation a neutron star will be spinning at about one thousand revolutions per second because it cannot lose its original angular speed. The same speeding-up occurs when a ballet dancer or skater draws in her arms from an outstretched position—but the effect is far more dramatic for a collapsing star.

All kinds of electric and magnetic

phenomena may come into play when a star with such an intense magnetic pole strength spins so fast. Shreds of material are likely to be torn from the surface of the star by electrical forces. This "atmosphere" may be caught up in the magnetic field and swirled round at the same rotation speed as the neutron star. A few hundred kilometers from the star, the material is likely to approach the velocity of light itself. (The theory of relativity states that no motion can exceed this velocity.) At this point gas must break away from the neutron star and fly into space as a stellar "wind."

STILL PUZZLES

Theoreticians puzzle over the strange phenomena that must take place near a pulsar. No truly satisfactory explanation of the "lighthouse" beam of radio emission yet exists. But there are many possibilities. Some theories hold that clusters of electrons—negative atomic particles—fill the rotating gas near the place where it must fly off the neutron star. Radiation emitted by these electrons would move at almost the speed of light (300,000 kilometers per second) and thus be concentrated into a narrow beam along its line of motion. Other theories suggest that streams of charged

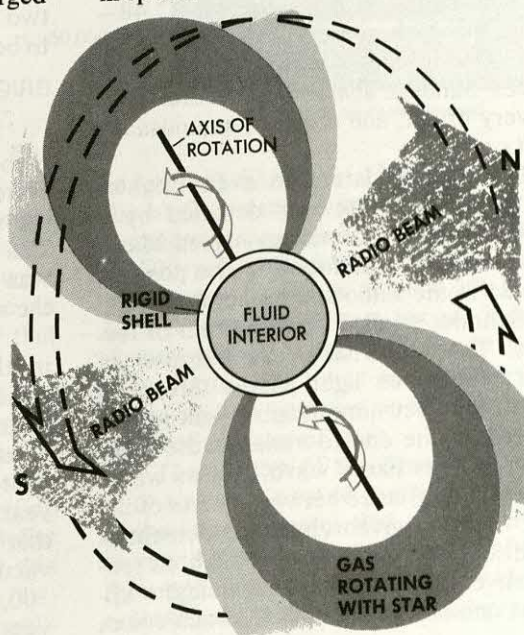
particles are ejected near the magnetic poles of the pulsar and so give rise to radio beams parallel to the axis of the magnetic field.

A neutron star, or pulsar, probably consists of a very rigid shell surrounding a fluid interior. Such a structure could explain the small changes in the period of the Crab pulsar. These may be caused by slight deformations of the shell, which in turn may be due to vibrations—"starquakes."

Whatever the final truth about pulsars may be, they have opened up a new chapter in astronomy. They have shed light on physical processes that could never be approached in any earthly laboratory. Compared to a neutron star, for example, the densest matter that could be found or made on earth would be almost a vacuum.

The discovery of pulsars has confirmed the existence of neutron stars. This brings up the questions: What about matter that is even too dense to become a stable neutron star? Would it continue to collapse gravitationally? Astronomers have framed a theoretical answer. It is the concept of the "black hole"—matter and energy so tightly compressed that nothing can escape. Scientists think they have discovered black holes in space.

Highly diagrammatic representation of a pulsar. The radio beams are emitted along the axes of the star's magnetic field (not shown). As the pulsar spins rapidly, the beams sweep around, causing a blinking, or pulsating, effect.



QUASARS

For several decades, it has been known that radio waves are emitted from various areas in the heavens. In the mid-1940's, scientists began a systematic study of radio sources—that is, areas emitting radio waves. The result was the development of a new branch of astronomy, known as radio astronomy, and a whole new family of instruments, including detecting devices called radio telescopes.

Using radio telescopes, astronomers located the positions of thousands of “radio stars”—that is, more or less sharply defined areas of radiation in the heavens. Sometimes radio stars could be identified with visible objects in the sky, but sometimes they could be detected only by radio telescope.

A STRANGE TYPE OF STAR

In 1960 an American astronomer, Alan R. Sandage, was engaged in photographing a portion of the sky that was emitting strong radio waves. When he studied his photographic plates, Sandage found what was apparently a star in the exact position of a known radio source. This particular radio source was numbered 3C 48—that is, the 48th source on the list of the Third Cambridge Catalogue of Radio Sources. Sandage also found that the object was very bright, and if a star, an unusually bright one.

Three years later, an even brighter object of this nature was detected by a Dutch-American astronomer named Maarten Schmidt. It was found in the position assigned to the radio source numbered 3C 273. Schmidt studied the spectrum of this object. The spectrum is the rainbow of colors seen when light passes through a prism. The spectrum of visible light ranges from red at one end to violet at the other end. Light acts like a wave, and its wavelength is the distance between crests of the wave. When a given chemical element is heated so that it glows, it emits light at specific wavelengths. These wavelengths appear as one or more bands of certain colors

of the spectrum. Furthermore, the bands of color are crossed by lines generated by individual emission and absorption lines of specific atoms. Therefore by studying the spectrum of light emitted by a star, astronomers can learn a great deal about the elements present in the star.

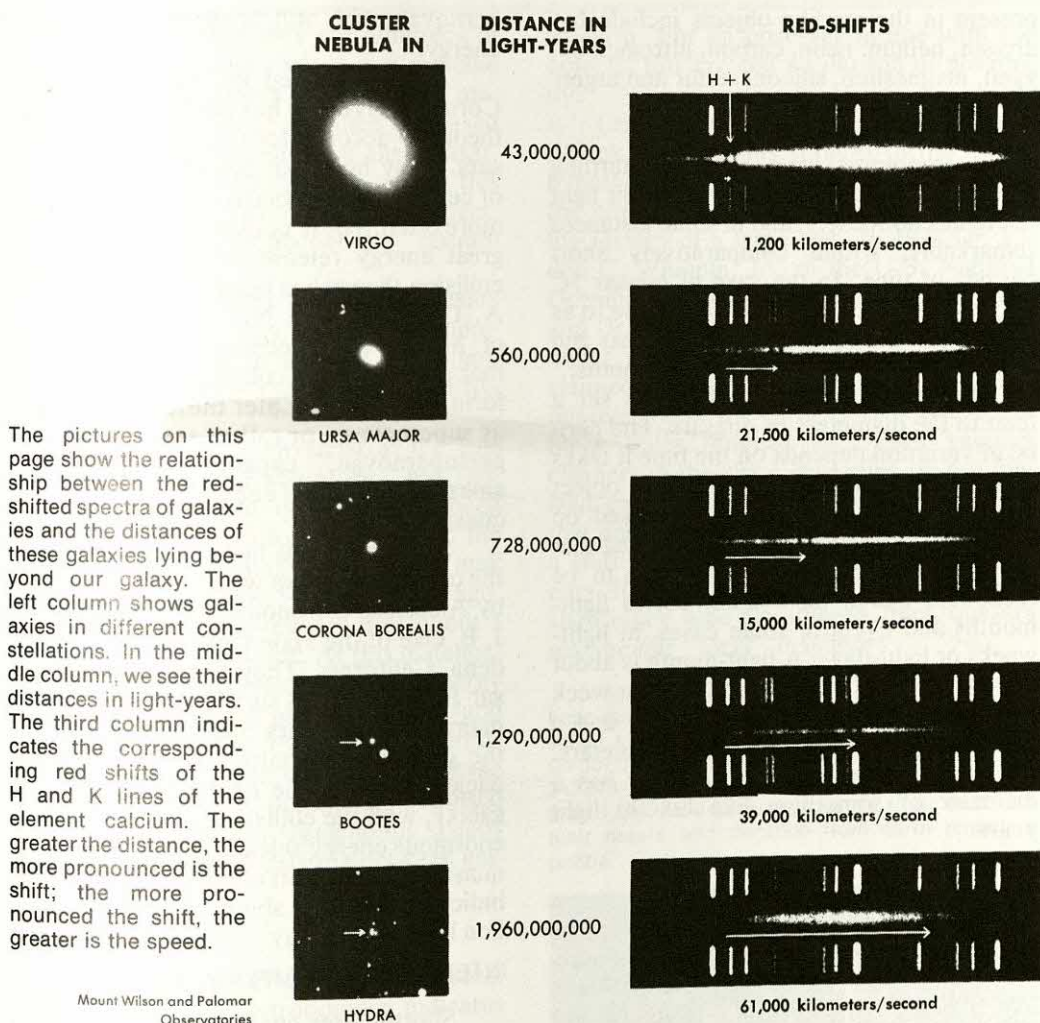
Study of the spectrum of 3C 273 and of the previously discovered object 3C 48 revealed that the spectra of these two starlike objects were not at all like the spectrum of a typical star. The spectral lines of the two objects were shifted toward the red end of the spectrum. When red shift occurs, all wavelengths are lengthened by the same amount, and the amount is expressed as a percentage increase over the normal wavelength. The red shift of 3C 273 amounted to 16 per cent; the red shift of 3C 48 to 37 per cent.

What is the significance of red shift? Most astronomers believe that red shift in the spectral lines of an astronomical object indicates that the object is receding from the earth. Conversely, wavelength shift to the violet end of the spectrum indicates that the object is moving toward the earth. The two starlike objects studied thus appeared to be receding from us.

BRIGHTER THAN ANY GALAXY

The American astronomer Edwin P. Hubble was the first to study the red shift systematically. He found that the greater the shift, the greater the speed of the star. The speed with which the star was receding was also proportional to the distance of the object from the earth.

The red shifts that had been noted in the starlike objects corresponding to radio sources 3C 48 and 3C 273 were very large. So great, in fact, that it appeared certain that they must be 1,000,000,000 light-years or more from the earth. A light-year is a measure of distance, the distance that light can travel in one year in a vacuum. The speed of light is roughly 300,000 kilometers per second, so one light-year is approximately 9,500,000,000,000



Mount Wilson and Palomar
Observatories

kilometers. Even at these fantastic distances the starlike objects were so bright that they must be emitting hitherto unheard of quantities of energy. Even if they were galaxies, the energy output would be so great that it could not be matched by that of any known galaxy.

Astronomers now began a search for other starlike objects that combined unusual brightness with strong radio emission and extreme red shifts. Over 1,000 have since been discovered, and it is estimated that up to 1,000,000 exist. The name "quasi-stellar objects" was first applied to them, but this has been shortened to "quasars."

No quasar has been found with a small-

er red shift than about 15 per cent. Usually the shift is much greater. In a very few instances it approaches the neighborhood of 300 per cent. Some of the quasars, at least, may be more than 4,000,000,000 light-years away. If quasars are indeed as distant as they seem, they would have to emit 50 to 100 times more energy than entire known galaxies consisting of hundreds of millions of stars.

The spectra of the quasars show lines of characteristic wavelengths. The strength of the lines indicates that they correspond to very hot gases, at temperatures of tens of thousands of degrees Celsius. The elements represented in the spectra and presumed

present in the starlike objects include hydrogen, helium, neon, carbon, nitrogen, oxygen, magnesium, silicon, sulfur and argon.

QUASARS ARE SMALL

Astronomers have noted a startling fact about quasars—namely, that their light fluctuates noticeably, and in some instances remarkably, within comparatively short periods of time. In the case of quasar 3C 355, the variations in brightness come to as much as 40 per cent in just a few weeks and to as much as 300 per cent in four months.

These variations in brightness set a limit to the diameters of quasars. The period of variation depends on the time it takes light to travel the diameter of the object emitting the light. Calculations based on variations in brightness have indicated that the diameters of most quasars are to be reckoned, not in light-years, but in light-months and even, in some cases, in light-weeks or light-days. A light-month is about 800,000,000,000 kilometers; a light-week about 181,000,000,000 kilometers; and a light-day about 26,000,000,000 kilometers. By way of comparison, our galaxy has a diameter of something like 80,000 light years.

BUT SO MUCH ENERGY?

The comparatively small size of quasars—if we assume that they are hundreds of millions of light-years away—makes their tremendous outpouring of energy all the more remarkable. How can we account for it? Various explanations about the nature of quasars have been offered.

According to one theory, proposed by George B. Field at the University of California, quasars are evolving galaxies. First, he suggests, an immense cloud of gas contracts. As the gaseous particles draw closer together, the gravitational force increases greatly. Finally, the entire mass collapses toward its center—a phenomenon called gravitational collapse. As a result, energy is emitted on a vast scale. In the course of time, stars are formed. As they build up larger atoms from smaller ones in the process of atomic fusion, more energy is released. Finally, these stars explode as su-

pernovae, with still another outpouring of energy.

A team headed by Thomas Gold of Cornell University has advanced a collision theory to account for the formation of quasars. They hold that the stars in the nuclei of certain galaxies begin to collide more and more often. These collisions account for the great energy release of quasars. Another collision theory has been offered by Stirling A. Colgate of the New Mexico Institute of Mining and Technology. He believes that as small stars collide, they combine to form larger stars. Later these stars explode as supernovae—or rather a number of “supersupernovae,” capable of emitting the amazing quantity of energy characteristic of quasars.

Support for this line of reasoning about the origin of quasars was presented in 1974 by American astronomers James Gunn and J. B. Oke of the Hale Observatory in Pasadena, California. They found that the quasar BL Lacertae is surrounded by a halo of many ordinary stars and may result from the collision of matter or small densely packed stars at the center of a spherical galaxy, with the collisions giving rise to the enormous energy release characteristic of a quasar. The red shift of quasar BL Lacertae indicates that it is about one hundred million light-years away.

INTERLOPERS COMPLICATE

Shortly after quasars were first recognized, the puzzle was further complicated when Sandage discovered that certain objects did not emit enough radiation to be classified as radio sources but were like quasars in other respects, including remarkably pronounced red shifts. These objects are sometimes called *interlopers*.

Quasar-like objects not known to be radio sources have now been found to be at least as numerous as quasars that emit strong radiation. It is not known what relationship, if any, exists between these two types of objects.

COULD QUASARS BE NEIGHBORS?

Thus far we have been assuming that quasars are at “cosmological distances”

from us—that is, that they are hundreds of millions of light-years distant and in some cases at the outer limits of the known universe. The estimates of distance are based on the degree of red shift of the quasar. However, not all astronomers accept this standard of quasar distance measurement, and thus some do not accept the “cosmological hypothesis.”

Some astronomers think that quasars may be comparatively nearby phenomena—nearby, that is, in terms of astronomical distances. This is called the “local hypothesis.”

James Terrell, a physicist at the Los Alamos Scientific Laboratory, has advanced the idea that quasars have been ejected from a gravitational collapse at the center of our own galaxy. This event may have taken place some 5 million years ago. The quasars, he believes, are now about 1 million light-years distant. At this distance, they would need to release far less energy to be strong emitters of radio waves and light radiation than if they were hundreds of millions of light-years away. As for the source of this energy, perhaps it is emitted as the quasars pass at tremendous speeds through clouds of gas existing between galaxies.

STILL A PUZZLE

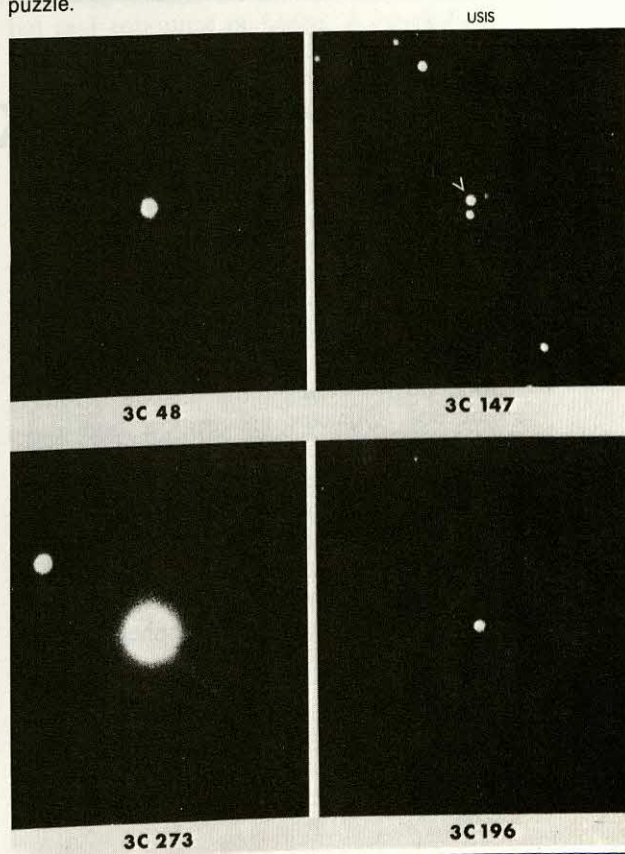
The debate about quasars continues. Research by some astronomers indicates that the brightness of a quasar decreases as its red shift increases. This supports the concept of red shift as an indicator of quasar distance and thus the cosmological hypothesis.

Other astronomers, however, point to several cases where photographs seem to show a quasar lying very close to or even connected to an ordinary galaxy, and yet the two objects have entirely different red shifts. These astronomers therefore claim that the degree of red shift is not an accurate indication of distance. In 1983 astronomers showed that the galaxy-quasar association could be merely an optical effect. They proposed that light from a distant quasar was bent by the gravity of a massive object in the halo of a galaxy closer

to earth. For an observer in line with the objects, this created an illusion of proximity.

Growing evidence for the existence of “black holes” lends support to the “local hypothesis.” A black hole is believed to be the completely collapsed last stage in the life history of a large star. In 1974 Dr. Eldon C. Whipple, Jr., of the U.S. National Oceanic and Atmospheric Administration and Dr. Thomas E. Holzer of the U.S. National Center for Atmospheric Research claimed that quasars are really black holes relatively near the earth. They explain that black holes function as “giant vacuum cleaners in the sky,” sweeping up hot ionized gases present throughout space, and that once these gases reach a certain speed a shift to the red end of the spectrum occurs. If quasar red shift is indeed caused by such black hole phenomena, then quasars may be our “local” neighbors.

Four quasars photographed using the Hale telescope at the Hale Observatories at Mt. Palomar, California. The first quasars identified—3 C 48 and 3 C 273—were discovered in the early 1960's. More than 1,000 have been discovered since then, but their nature and distance from earth remain a puzzle.





An artist's conception of a black hole.

H. K. Wimmer

BLACK HOLES

by Mort La Brecque

Black holes are the most fascinating and mysterious objects in the heavens.

In the 1960's, the most important discoveries in astronomy were pulsars and quasars. Pulsars are regularly pulsating radio and (in at least one instance) optical sources. Quasars are optical and radio sources of enormous intensity, apparently at great distances from the earth.

The detection of pulsars and quasars, made possible largely through advances in radio astronomy, led to the search in the 1970's for a new class of objects that may be the most bizarre physical phenomena in the universe.

These phenomena are called black holes. They are so called because they give off no light and act like stellar vacuum cleaners, sucking in matter and energy from space.

Black holes, which are very small, are proposed by astrophysicists as the last stage in the life history of very large stars. Collapsed by the force of their own gravity, black holes are deduced by scientists from Albert Einstein's general theory of relativity. Einstein's theory drastically revised Newton's concept of gravitation.

If a black hole is detected in outer space—and one is believed to have been

discovered—the event will be significant for physics as well as for astronomy. Classical physics cannot account for a black hole. If one exists, general relativity will be virtually confirmed.

A PRODUCT OF AGE

Extraordinary as they are, black holes are merely products of a universal phenomenon: physical aging. After thousands of millions of years, stars burn up their hydrogen fuel and begin to cool and contract. As their dimensions decrease, their gravitational forces increase. Eventually, all stars collapse under their own gravity.

Like living things, stars resist gravitational collapse and death. Dying stars produce internal pressures to fight the awesome force of gravity, but nature plays a devious “trick” on them. Energy is equivalent to mass. As the stars produce more internal energy, they increase their effective mass and gravitational attraction. Thus, the ultimate fate of a particular star depends on its mass. “The bigger they are, the harder they fall” applies to stars as well as to prizefighters. The more massive the star, the stronger its internal energy, the larger its gravitational attraction, and the greater its collapse.

WHITE DWARFS AND NEUTRON STARS

Small stars, such as our sun, collapse to objects called white dwarfs. About the size of the planet earth, white dwarfs resist further collapse with internal pressure caused by electrons spinning at near the velocity of light—300,000 kilometers a second. White dwarfs are very dense objects—a cubic centimeter weighs several tons.

But white dwarfs are lightweights compared to neutron stars. These objects are the evolutionary end products of larger stars, from 1.4 to 2 times as large as the sun. Electrons cannot resist the greater gravitational collapse of such stars and are pushed into atomic nuclei, where they combine with protons to form uncharged, tightly packed neutrons. Neutron stars are only a few kilometers in diameter. They weigh about one million tons per cubic cen-

timeter. They can resist further collapse only by “invoking” the strongest force in nature—appropriately called the strong force—which binds atomic nuclei.

The strong force halts the imploding matter so abruptly—in a tenth of a second—that the collapsed stellar cores act as charges to set off supernova explosions in the stars’ outer portions. Such celestial fireworks, observed by Chinese astronomers in July 1054, produced the Crab nebula, a cloud of gas that still writhes and glows 6,000 light-years from earth. A light-year is a measure of distance, equivalent to the distance that light can travel in one year, or about 9,600,000,000,000 kilometers.

LARGE STARS BECOME BLACK HOLES

What happens to a dying star that is more than twice as large as the sun? Even the strong force cannot halt its infalling momentum, and it collapses completely, beyond the neutron-star stage, to an even smaller, denser object, the black hole. Complete collapse does not mean that the black hole vanishes from the universe. The structure of space-time, as described by Einstein, precludes an infinite collapse and produces instead an immaterial, invisible but real curvature of space. A black hole can be compared to a hefty invisible man who sits on a couch. He cannot be seen, but his weight creates a depression in the seat.

Black holes are nothing new to theoretical physicists. They were first proposed in 1939 by J. Robert Oppenheimer and Hartland S. Snyder as a consequence of general relativity, but there was no way known of detecting them at that time.

However, with the recent development of radio astronomy and the detection of inexplicable radio signals from deep space, black holes have become a subject of great interest to astronomy. It is believed that these theoretical objects could play a role in such extraordinary energy phenomena as quasars and pulsars. Black holes and neutron stars are the only objects known to physics that are sufficiently compact and massive to fulfill astronomical observations of those very strong emitters of radiation.

PROPERTIES OF BLACK HOLES

With little equipment, physicists have developed a fairly comprehensive description of black holes. Unlike every other physical object, black holes have neither size nor shape in the conventional sense, according to Drs. John Wheeler and Remo Ruffini of Princeton University. But they function within a diameter of about 15 kilometers; they have masses ranging from that of the sun to a hundred million times as much; and they act like vortices. Any stray matter or energy that passes too close to a black hole—within a critical distance called its horizon—will be irresistibly drawn into the vortex which is the black hole. Violent tidal forces within the black hole stretch the matter in one direction and squash it in another, until it literally decomposes to become part of the black hole's curved space.

Other black-hole properties are even stranger. Space and time exchange their characteristics inside the completely collapsed star. Under normal conditions, an object maintains its size but is subject to physical aging. Inside the black hole, it doesn't age, but continuously becomes smaller. Observers at a safe distance from the black hole could not actually see it, because light, like other forms of energy, is vulnerable to a black hole's suction. As light is drawn in, it shifts infinitely to the red end of the color spectrum, rendering the black hole black and therefore invisible. If black holes were somehow visible, observers would see these stars as they appeared just before collapse, even if the collapse had occurred thousands of millions of years earlier. This is because as soon as the star becomes a black hole, it is frozen in time with reference to observers outside it. "All signals and all information from the later phases of collapse never escape; they are caught up in the collapse of the [space-time] geometry itself," according to Drs. Wheeler and Ruffini.

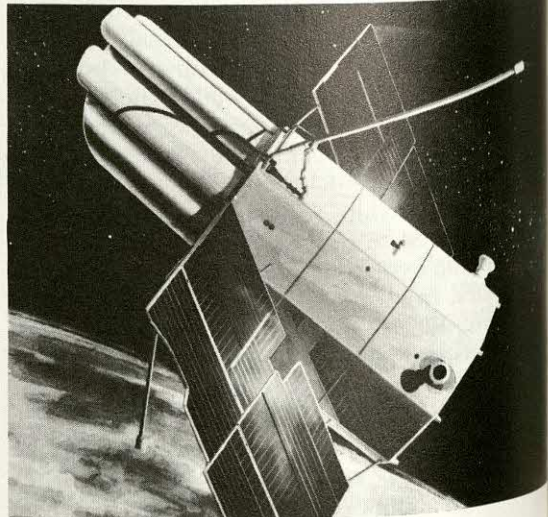
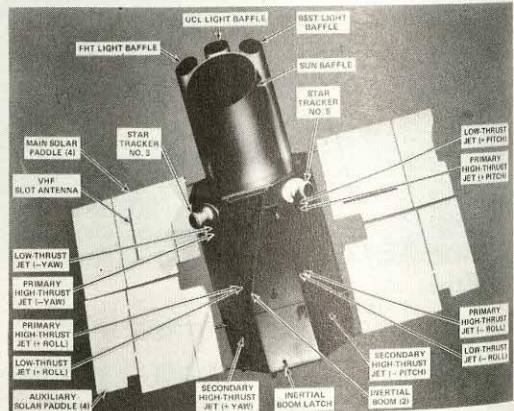
HOW MANY ARE THERE

How many black holes are there in the universe? According to A. G. W. Cameron of Yeshiva University, the universe may be

teeming with them. Cosmological theory predicts that the universe contains a precise amount of matter. But astronomers have deduced from their observations that there is not nearly enough matter to satisfy the predictions. Observed matter is less than predicted matter by a very considerable amount. Dr. Cameron suggests that the missing matter may have been swallowed up by large numbers of black holes.

The chemical history of the universe indicates that the first stars to be formed were very massive and would be expected to evolve into black holes. It is not known Copernicus, code name for the latest orbiting astronomical observatories, is specially equipped with X-ray and ultraviolet telescopes to study star phenomena. Below: diagram showing some of Copernicus' equipment and artist's idea of Copernicus in orbit.

Grumman Corp.



with certainty that all large stars inevitably develop into black holes. Scientists have not shown that highly asymmetric stars—stars not nearly as symmetric as perfect spheres—are subject to this fate. Minor asymmetries in shape, however, will not save a large star, according to the Soviet physicist Y. B. Zeldovich and the English team of Steven Hawking, Roger Penrose and Robert Geroch.

DETECTING BLACK HOLES

One way of detecting black holes is through the gravity waves they emit at the time they collapse. Any stellar mass of asymmetrical shape gives off gravitational radiation. But only gravitational collapse, involving very large masses and very rapid increases in radiation, is expected to provide an obviously detectable source. Joseph Weber of the University of Maryland, a pioneer in the field of gravitational radiation, has detected many events that indicate the large-scale destruction of matter in the universe through gravitational collapse. His equipment consists of instrumented aluminum antennas suspended by wires inside

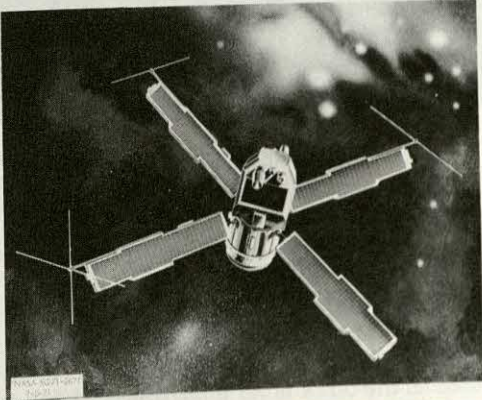
shielded chambers. This equipment is capable of detecting black holes but, unfortunately, not with precision.

However, a way around this problem has been proposed, which takes advantage of the present state of astronomical technology. In fact, astronomers believe they have already located a black hole in space.

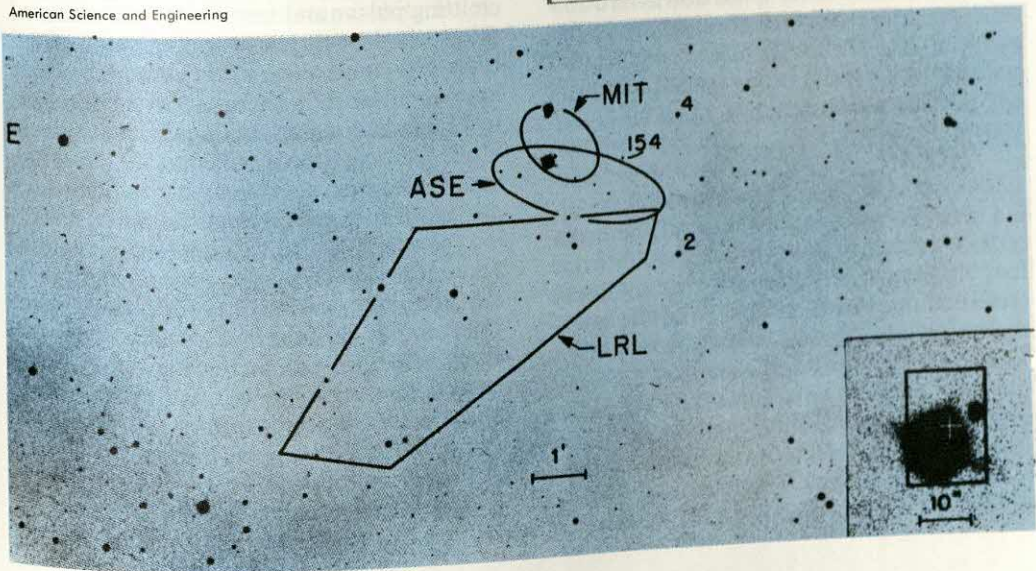
When stray matter, such as is contained in interstellar gas clouds, drifts past a black hole's horizon, it funnels down into the collapsed star. Professors Zeldovich and I. D. Novikov proposed that the matter is then heated by compression and radiates energy.

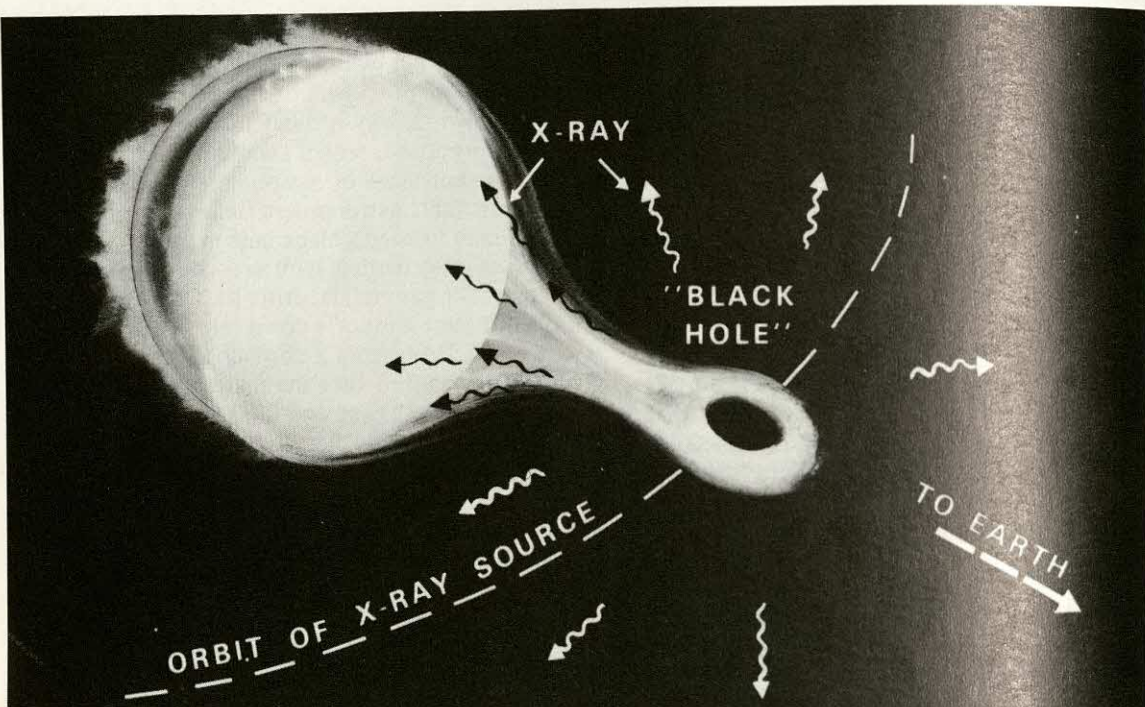
Black holes, however, would not pass through gas clouds very often. A better chance of locating one occurs when the black hole belongs to a binary-star system (two stars that revolve around each other)

NASA



The Uhuru satellite, right, was the first satellite designed for X-ray astronomy. It located the X-ray source Cygnus X-1, whose location is shown below in the area of overlap between the error boxes labeled MIT and ASE. Detail is shown at lower right.





NASA

Artist's conception of what may be the arrangement of a binary star system, in which one star is visible and the other is a black hole in orbit around it. Astronomers have now tied the binary system HDE 226868 to the X-ray source Cygnus X-1 and have found evidence of the visible star's gas clouds swirling into Cygnus X-1, which has been identified as a black hole.

and is drawing in matter constantly. Matter from thermonuclear explosions on the nearby star's surface will be pulled into orbit around the black hole, spiral downward and produce large emissions of radiation. In the early 1970's Drs. Wheeler and Ruffini stated that such emissions could be detected by X-ray astronomy.

A POSSIBLE BLACK HOLE

In 1970, a United States artificial satellite, the Uhuru, was launched off the coast of East Africa. (Uhuru is the Swahili word for "freedom.") Its purpose was to detect stellar sources of X rays, free of the interference of the earth's atmosphere. Uhuru has found over 100 stars that give off X-ray pulses. The three strongest are: Cygnus X-1 in the constellation Cygnus; Centaurus X-3 in the constellation Centaurus; and Lupus X-1 in Lupus. The first studies were of Cygnus X-1 and Centaurus X-3.

Cygnus X-1 pulses very rapidly. Its rate of pulsation is second only to NP-0532 in the Crab nebula, an X-ray- and radiowave-emitting pulsar and very likely a remnant of the A.D. 1054 supernova explosion. Centaurus X-3 pulses far more slowly than any known pulsar, but unexpectedly discharges as much energy as the Crab pulsar does. Unlike conventional pulsars, Cygnus X-1 and Centaurus X-3 emit no radio waves and have no detectable gas clouds.

These X-ray sources must be rotating, gravitationally-collapsed objects, but they are certainly not ordinary neutron stars. Three other possibilities remain: they are peculiar neutron stars, or they are an unknown class of objects, or they are black holes producing great quantities of radiation when matter funnels into them.

The argument for black holes was greatly strengthened late in 1971 when Uhuru data indicated that Cygnus X-1 lies

very close to a massive, old supergiant star. This suggests that it is a member of a binary-star system. Observing the supergiant optically, astronomers found that it is circling an unseen object once every 5.6 days. They also found what they considered to be a cycle of X-ray emissions. Uhuru subsequently began extended observations and verified the length of the cycle.

In 1974 astronomers in England and the United States officially designated Cygnus X-1 as a black hole. In 1983 U.S. and Canadian astronomers told of another candidate: an invisible X-ray source in the Large Magellanic Cloud that is circled in less than two days by a faint blue star.

MODEL OF THE UNIVERSE?

If astronomy has finally responded to physics' proposal of black holes, physicists are thanking astronomers in a time-honored fashion of science: by setting them a new and difficult task. Einstein's general theory of relativity is mathematically time-symmetric: the theory indicates that relativistic processes operate in two opposite directions. It is like a movie that can be shown running backward as well as forward. This means that while black holes implode and absorb matter and energy, there must also be stars that simultaneously explode and emit matter and energy somewhere in the universe. Color-conscious physicists have named these theoretical stars white holes. At present, nobody knows how white holes express themselves in space, or what role they play in stellar energy phenomena.

Black holes and white holes may have an importance greater than that of all stars; galaxies, and clusters because they may offer a model of the life, death, and rebirth of the universe itself. According to one cosmological theory currently favored by many physicists and astronomers, the universe has a life cycle like all the physical and biological objects within it. The cycle begins with an explosion of incredible magnitude—the big bang—which sends matter and energy flying apart in all directions. At first, the objects within the universe recede from one another at great speed, as if they were spots on an expanding balloon. The



Sovfoto

This 1929 photograph of the Tunguska region of Siberia shows trees charred by a mysterious explosion in 1908—an explosion many believe was caused by a comet or meteorite impact but some now speculate may have been caused by a "mini" black hole colliding with the earth.

force of gravitation gradually slows down their expansion. Eventually, their mutual gravitational attraction causes the stars to reverse direction and come back toward one another at an increasing speed, somewhat the way the spots will rush together when air is let out of the balloon. Finally, all the matter in the universe reaches the same point simultaneously, and another big bang occurs to begin another cycle of the universe. Some astronomers question this theory, however. Some believe that there is too little matter in the universe to create the gravitational attraction necessary for the contraction phase of the cycle, and that therefore the universe will expand forever. Still others take an opposite view, believing that the force of gravity will draw all matter together until it collapses.

What do black holes and white holes have to do with theories of the universe? It is believed that the physical processes that occur when a large star suffers gravitational collapse and becomes a black hole are the same processes that would be involved in the collapse of the universe. Similarly, whatever happens inside a white hole would be identical with the big bang. Many scientists believe that the discovery and study of these objects will increase our understanding of the universe.

ROCKETS

by Willy Ley

The statement has often been made that the space age began on October 4, 1957, when the Soviet Union put the first artificial satellite—Sputnik 1—into orbit around the earth. Actually this event was the culmination of many years of thought and work. Before an artificial satellite could be orbited, rockets large enough and fast enough to do the job had to be designed and tested. Before these rockets could be built, the science of physics had to find the natural laws that applied to rockets. The science of engineering had to develop the necessary materials and techniques.

The first, faltering steps that led to today's achievements were taken less than a hundred years ago. But interest in flights to heavenly bodies in space goes back many centuries. For example, we find two tales of trips to the moon in the writings of Lucian of Samosata, a Greek writer of the second century. In his *True History*, a ship is blown all the way to the moon by a storm. The tale called Ikaromenippus tells of a man who flies to the moon, using wings made from the feathers of large birds.

In all the earlier writings on the subject it was assumed that the earth's atmosphere reached all the way to the moon and other heavenly bodies. The problem, it seemed, was to perfect a craft that could fly through the air. We now know that the air, or atmosphere, is a film that extends only a comparatively short distance above the ground. Ninety-nine per cent of this film lies within 32 kilometers of the earth's surface. At an altitude of 160 kilometers, the molecules and ions (charged particles) making up the atmosphere are few and far between. Because of this, airborne devices such as balloons, dirigibles, and airplanes have not been stepping-stones to space, since none of them can stay aloft without a supporting atmosphere.

Yet, curiously enough, a device that can bridge the gulf of space between the earth and other heavenly bodies has been



Estes Industries

Here, in miniature, the excitement of the lift-off of a full-scale rocket is being duplicated by model rocketeers.

known to mankind for centuries. This device is the rocket, which was used by the Chinese in warfare as early as the thirteenth century. The device is based upon Sir Isaac Newton's third law of motion, which states that for every action there must be a reaction, equal in force but opposite in direction. This law explains the recoil of a gun or rifle. At the moment of firing, there is a backward thrust (recoil) corresponding to the forward thrust of this weapon; in the case of a big gun, this recoil is very formidable. In a rocket, the firing of the fuel causes gases to be expelled from the rear. The recoil causes the rocket to dart forward in the opposite direction.

As long as the rocket fuel in a large modern rocket burns and shoots out thousands of millions of molecular bullets—the combustion gases—there will be recoil that will push the rocket ahead. The motion is virtually independent of the air surrounding the rocket. In fact, the device would function even more efficiently in an absolute vacuum. For one thing, there would be no air to resist the rocket's forward motion. Even more important, there would be no air resistance to impede the exhaust. Hence the exhaust velocity would be greater than if the rocket were traveling through the atmosphere; the thrust would be greater. The thrust of the rocket depends on only two factors. The first is the quantity of combustion gases produced (which, of course, corresponds to the amount of fuel being burned). The second factor is the speed with which these gases are ejected.

Jet-propelled aircraft also move by recoil, but such craft could never be used for space flight. The reason is that a jet engine depends upon the oxygen in the air for the combustion of its fuel. In the typical rocket, the oxygen needed to burn the fuel is carried along, either in a separate tank (in liquid-fuel rockets) or in the form of an oxygen-rich chemical that is mixed with the fuel (in solid-fuel rockets).

EARLY RESEARCH

The first man to realize the possibilities of the rocket in space flight was a German inventor, Hermann Ganswindt, who began to lecture on rockets and space flight in the last decade of the nineteenth century. A contemporary of his—a Russian school-teacher named Konstantin Eduardovitch Tsiolkovsky—also called attention to the rocket as a device capable of penetrating the atmosphere and passing into outer space.

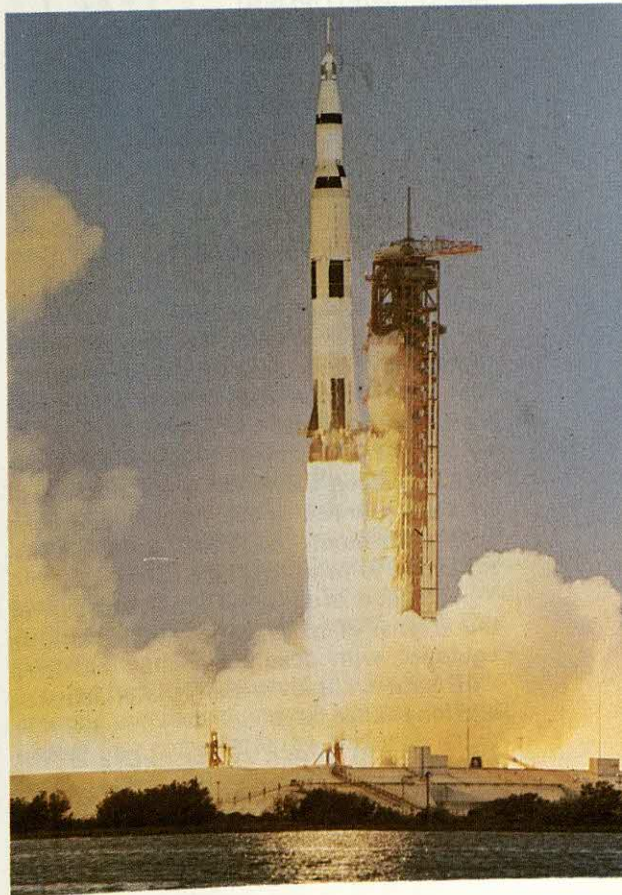
Neither Ganswindt nor Tsiolkovsky created much of a stir. For one thing, the only rockets in existence at the time were of the so-called black-powder type. Black powder—that is, black gunpowder, made from charcoal, saltpeter (potassium nitrate), and sulfur—is about the weakest fuel that could be used to propel a rocket. Besides,

most people thought of rockets as devices used for fireworks. Hence the idea of a rocket soaring far out into space struck them as comical. Ganswindt and Tsiolkovsky were theorists; neither performed experiments with the idea of improving rocket performance.

The first man to do serious research in rocket development was an American, Professor Robert H. Goddard, of Clark University in Worcester, Massachusetts. He systematically investigated various kinds of gunpowder to find which had the greatest exhaust velocity when used in a rocket. He also made a mathematical analysis of rocket motion. Goddard's findings were presented

Saturn V rockets launched Apollo spacecraft from Kennedy Space Center in the 1970's. Really three rockets in one, Saturn is able to achieve the great speeds needed to put a spacecraft into orbit.

UPI



in 1919 in a Smithsonian report entitled *A Method of Reaching Extreme Altitudes*.

Four years later, a book called *The Rocket to Interplanetary Space* (*Die Rakete zu den Planetenräumen*), by Professor Hermann Oberth, was published in Germany. It strongly advocated the use of liquid fuels, discussed the possible construction of liquid-fuel rockets, and considered the problems of orbiting the earth with such devices. Oberth's book led to the formation of the Society for Space Travel (*Verein für Raumschiffahrt*) in 1927, with the goal of actually building the rockets contemplated by Oberth. In the meantime, Professor Goddard had also become interested in liquid-fuel rockets, and he constructed a number of them. On March 16, 1926, one of these craft made a short flight from a farm near Auburn, Massachusetts. It was the first liquid-fuel-rocket flight in history.

RAPID DEVELOPMENT

Today the rocket has reached a stage of development that would have seemed fantastic even a generation or two ago. The foremost achievement technologically has been the development of advanced rocketry for the exploration of space. Every flight into space begins with a rocket launching. Many artificial satellites have been placed in orbit and are invaluable for communications, meteorological, and earth observation purposes. Unmanned space probes have been sent to Mercury, Venus, Mars, and Jupiter with precise maneuvers for favorable observation sites carried out by small rocket firings. And, in what has probably been the most spectacular use of rockets, manned space flights have been sent to the moon and to establish and transfer crews to orbiting space stations.

Rockets are also an important part of the arsenal of many nations. They can be equipped with deadly weapons and fired with great accuracy at short-, intermediate-, and long-range targets.

TWO MAJOR TYPES OF ROCKETS

It is now time to take a closer look at the rocket, the very symbol of the space age. All rockets in actual use at the present

time belong to one or the other of two groups: liquid-fuel and solid-fuel. Thus far it is the liquid-fuel rocket that has been the workhorse of the space program. The solid-fuel rocket has served mainly as a military weapon. However, it has also played a minor part in space exploration and could conceivably play a larger part later on. Other rocket types are under development, including the hybrid rocket, the atomic rocket, and the ion-drive rocket. We shall consider each of these types in turn.

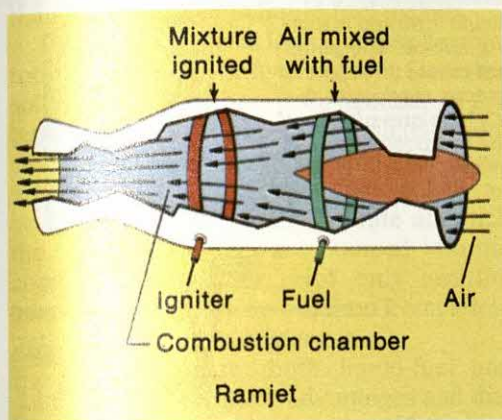
LIQUID-FUEL ROCKET

The typical liquid-fuel rocket used in space flight consists of the following parts from top to bottom, as it stands on the firing pad: the nose cone, the instrument section, the fuel-tank section, and the rocket engine. Actually, as we shall see, there are always two or more rockets in series in space flights, each having its own instrument section, fuel-tank section, and rocket engine. There is one nose cone for the combined rocket.

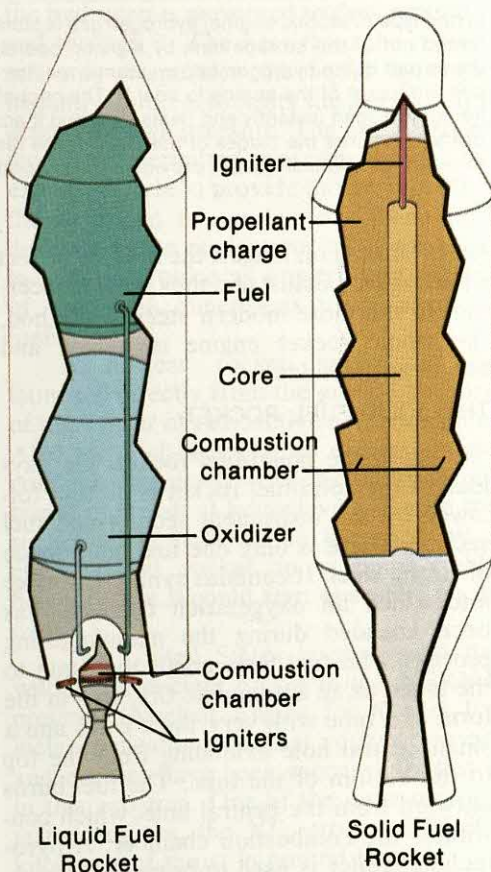
The nose cone holds what engineers call the "payload." It may consist of a package of scientific instruments, in the case of a research rocket. It may be a manned space capsule which is to go in orbit around the earth or the moon. In a ballistic missile, the payload is the nuclear warhead.

The instrument section (also called the guidance section) contains the instruments that guide and regulate the flight of the rocket and also transmit information about the rocket's flight to the ground.

The fuel-tank section is the bulkiest part of a rocket. One tank in this section holds the fuel. In another tank is the oxidizer, which causes combustion when it reacts chemically with the fuel. In most liquid-fuel rockets the fuel is either alcohol or refined kerosene (called RP-1, "RP" standing for "rocket propellant"); the oxidizer is liquid oxygen. In rockets of this type, ignition must be provided to bring about combustion. There are various types of igniters. Pyrotechnic powder squibs (really fireworks), set off by electricity, may serve the purpose. Ignition may also be provided by



Both jet-propelled planes and rockets are based on the recoil principle, described in the text. However, in the jet plane, such as the ramjet shown above, air (which contains oxygen) is drawn in and mixed with injected fuel and the mixture is then ignited. A rocket, on the other hand, carries its own oxygen supply in the form of an oxidizer. In the liquid-fuel rocket (near right), the fuel is separate from the oxidizer. Fuel is mixed with the oxidizer and ignited in the combustion chamber. In the solid-fuel rocket (far right), the oxidizer is contained in the fuel itself. After the fuel is ignited, it burns outward from a central core. The fuel tank itself thus becomes the combustion chamber.



plugs that are made to glow as electric current courses through them.

Certain fuels and oxidizers require no igniter when used in combination, since they start burning on contact. These are known as hypergolic propellants. Thus the hypergolic liquid aniline will burst into flame if it comes in contact with an oxidizer such as nitric acid (a compound of oxygen, hydrogen and nitrogen: HNO_3). In the most frequently used hypergolic combination, a chemical called unsymmetrical dimethyl hydrazine (abbreviated UDH and sometimes called dimazine) is the fuel; nitric acid is the oxidizer.

A fuel consisting of hydrogen and oxygen is now used for upper-stage rockets. These rockets are considered to be of the liquid-fuel variety, for the hydrogen and oxygen are used in liquid form.

Below the fuel-tank section is the rocket engine, consisting of the rocket motor

and several auxiliary devices such as pumps, turbines, and rudders. The rocket motor has two basic parts. One of these is the combustion chamber, where the fuel and oxidizer come together and are burned. The other is the exhaust nozzle, which is shaped in such a way that the exhaust blast resulting from combustion leaves with the highest possible velocity. Pumps force the two liquids—fuel and oxidizer—into the combustion chamber. These pumps are of the centrifugal variety and are driven by a turbine. In modern liquid-fuel rockets, the turbine is driven by the same fuel that powers the rocket.

An important element of a rocket is the steering mechanism, which carries out the orders that come from the guidance equipment. There are two ways of steering a large liquid-fuel rocket. One method employs rudders set in the exhaust blast and made of graphite to withstand the intense

In one type of atomic engine, hydrogen gas is stored in liquid form and is then forced out of the storage tank by a pump operated by a turbine. The pump drives part of the hydrogen into an atomic reactor; the rest passes through the cooling jacket of the engine to cool it. The heat of the reactor evaporates the liquid hydrogen instantly and heats it so that it acquires a high pressure. Part of the gas turns the blades of the turbine; the rest of the hydrogen gas goes through the exhaust nozzle, providing thrust for the rocket.

heat. Turning on hinges, the rudders deflect a part of the exhaust and thus provide steering. In the more modern steering method, the whole rocket engine is hinged and serves as the rudder.

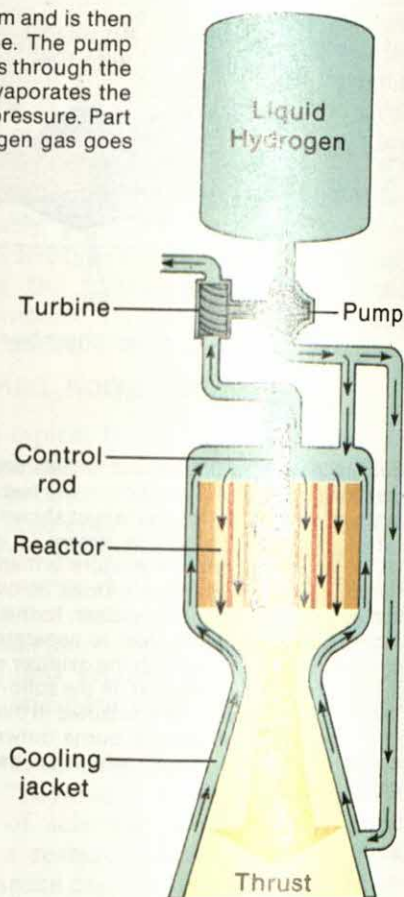
THE SOLID-FUEL ROCKET

As in the liquid-fuel rocket, the payload of the solid-fuel rocket is on top, followed by the instrument section and fuel section. There is only one fuel tank, made of strong steel. It contains synthetic rubber into which an oxygen-rich chemical has been kneaded during the manufacturing process. The fuel feels about the same to the touch as an automobile tire. It is in the form of a tube with very thick walls and a small central hole extending from the top to the bottom of the tube. The fuel burns outward from the central hole, which constitutes the combustion chamber. A pyrotechnic igniter is used to bring about combustion. This igniter consists of an electric wire surrounded by a powder charge; when the wire is heated, it sets off the charge.

In a solid-fuel rocket, the "fuel tank" and the "combustion chamber" are one and the same. The steel tube containing the fuel is also the motor, and that is the name given to it. The exhaust nozzle is attached (usually welded) to the lower end. The rocket can be steered by setting rudders in its exhaust blast, as in the liquid-fuel motor. Another method consists of injecting a high-pressure inert gas into the exhaust nozzle, thus deflecting the exhaust blast.

A COMPARISON

At first glance, the solid-fuel rocket would seem to offer certain definite advantages. For one thing, it is a fairly simple device, consisting basically of a large tube of high-grade steel. Hence it is much cheaper to construct and maintain than the liquid-



fuel rocket. The latter is a complicated device, requiring even more careful servicing than a passenger-carrying airliner. Another point in favor of the solid-fuel rocket is that once the synthetic-rubber fuel is in place, it can be readied for firing in a short time. The firing of a liquid-fuel rocket is a complicated procedure. In the countdown before the firing of such a rocket, many things can, and sometimes do, go wrong.

However, solid-fuel rockets also have certain disadvantages. Since the steel tube holding the fuel charge has to be strong, it is unusually heavy. In the liquid-fuel rocket, combustion does not take place in the fuel tanks that make up most of the rocket's bulk, so that the "skin," or covering, can be comparatively thin. Hence the structural weight of the liquid-fuel rocket can be kept down. As for the synthetic-rubber fuel of the solid-fuel rocket, it costs much more

than the fuels used in liquid-fuel rockets.

The solid-fuel rocket is far less flexible, too, than the liquid-fuel variety. Once the solid-fuel charge has been ignited, it will burn until it has been completely consumed; its thrust cannot be modified. However, the thrust of a liquid-fuel rocket can be adjusted by changing the rate at which the fuel and oxidizer are pumped into the combustion chamber. Not only can the burning of the fuel be stopped completely but it can also be started again.

To summarize, both liquid-fuel and solid-fuel rockets have advantages and disadvantages. Probably because of its high structural weight, its lack of flexibility, and its costly fuel, the solid-fuel rocket will never be so important a factor in space flight as the liquid-fuel type. It has served in rocket launches, but only on a small scale. Probably its chief use in space flight will be in combination with liquid-fuel rockets. The Titan III C combines solid and liquid fuels.

OTHER TYPES OF ROCKETS

An interesting rocket now under development is the hybrid. It is so called because, though its fuel is solid, its oxidizer is not. In this case, the oxidizer is gaseous oxygen. There is a spherical pressure tank for oxygen and a tubular fuel charge with a central hole. When the rocket is to be fired, oxygen gas is made to flow through the center hole of the fuel charge, and ignition is provided at the same moment. Hybrid rockets combine the flexibility of the liquid-fuel rocket—since the flow of oxygen gas can be regulated and stopped—with the simplicity of the solid-fuel rocket. But the structural weight is even higher proportionately than that of a steel-cased solid-fuel rocket. The chances are, therefore, that hybrid rockets will be used only as auxiliary devices.

Another possible rocket of the future is the nuclear rocket. Here the "rocket motor" is an atomic reactor designed to liberate large quantities of heat. Above this reactor is a tank holding liquid hydrogen. There are also various pumps and auxiliary devices. When the nuclear reactor is hot,

the hydrogen is permitted to flow around it and through it. The heat from the reactor not only evaporates the liquid hydrogen instantly, but it also heats the gas so that it acquires a high pressure. The hot hydrogen gas is then permitted to escape through the exhaust nozzle to provide thrust. Note that the hydrogen is *not* ignited. When fully developed, the nuclear rocket is expected to be about twice as powerful as rockets of the same dimensions using chemical fuels.

A nuclear rocket might not be launched directly from the ground because of the danger of radioactive contamination. Also the high-temperature exhaust might combine chemically with the oxygen contained in the air at lower levels. Hence the rocket might have to be launched by a chemical-fuel rocket to a considerable height before it could start operating on its own.

The United States is at present developing a nuclear rocket in the NERVA program (*Nuclear Engine for Rocket Vehicle Application*). At least a dozen reactors and engines have been successfully tested in this program. One of the most recently tested engines, the XE, produced 25,000 kilograms of thrust in ground tests.

An ion-drive rocket is also under development. In this device, electrically neutral atoms or molecules are converted into ions carrying an electric charge. The ions are then accelerated by means of electrical fields to form a very rapid exhaust. Because the quantity of ions that can be produced in the device is small, it is desirable that they should be heavy in order to achieve maximum thrust. This means that they should be ions of a heavy metal. Unfortunately most such metals happen to be naturally radioactive, and this would bring about complications, including a danger factor. The comparatively few heavy metals that are not radioactive are generally rare. In fact, there are only two metals that qualify in every respect for the proposed ion-drive rocket: cesium and mercury. Both have been tested in the laboratory and by means of vertical rocket shots going beyond the atmosphere.

ACHIEVING SPACE FLIGHT

In order to be used effectively in space missions, a rocket must have a satisfactory *mass ratio*. It must achieve *circular velocity* if its payload is to go into orbit around the earth. It must achieve *escape velocity* if its payload is to escape entirely from the earth's gravitational attraction.

Mass ratio. To explain what this term means, we should first point out that the weight of any rocket, ready for takeoff, is the sum of three units. The first of these is the payload, which is the reason for the flight. The second unit is the fuel load. The third is the dead weight, or the dry weight, as the British call it. It is equivalent to the total weight of the rocket structure, including the rocket engine, fuel pumps and the like, but not including the payload or the fuel. Payload, fuel load, dead weight—these together make up the takeoff mass. After the fuel has been burned, payload and empty rocket are the "remaining mass." If you divide the takeoff mass by the remaining mass, you obtain the mass ratio.

Let us say, by way of example, that the takeoff mass is 45 metric tons and the remaining mass is 15 metric tons. In this case the mass ratio is 3:1. Suppose you could build a rocket with the same takeoff mass but with a lower structural weight, so that your fuel load would be a higher percentage of the total. You would then have a rocket that would burn longer; it would reach a higher velocity and would have a longer range. A goal of good rocket engineering therefore is to make the mass ratio large. This is much the same thing as saying that the dead weight must be kept comparatively low. No rocket intended for space flight can be successful unless this objective is met.

Circular velocity and escape velocity. The circular velocity is that which is required for a rocket to put a payload in orbit, without further expenditure of fuel, around a planet, such as the earth. The escape velocity is that required for the payload to escape from the gravitational attraction of that planet. It can be determined by mul-

tiplying the circular velocity by the square root of 2, or approximately 1.41. The figures for circular velocity and escape velocity differ for each planet because they depend on the planet's mass. In the case of the earth, the circular velocity is approximately 8 kilometers per second, the escape velocity, 11 kilometers per second. If we lived on the smaller planet Mars, the velocities would be lower.

Actually, no single rocket fired from the earth has ever achieved circular velocity, let alone escape velocity. Even if the mass ratio were 7.4:1, the rocket's velocity would not exceed 4,900 meters per second—that is, only about 5 kilometers per second. But any rocket engineer will tell you that even with the best methods of design and construction, it is impossible to build a rocket with the mass ratio 7.4:1.

The step principle. Circular velocity and escape velocity are achieved by applying the step principle. By this we mean that the payload of one rocket is another rocket—a much smaller rocket, of course, since it must fit in the first one. When the fuel of the first step, or stage, has been used up, the second stage takes over. This stage is moving rapidly before it even begins to operate on its own; hence it attains a much greater speed than the first stage. We could add a third or fourth or fifth stage.

Let us assume that the rocket of the first stage has a mass ratio of 2.72:1 and reaches a velocity of 2,400 meters per second. The second stage, which was the payload of the first stage and also has a mass ratio of 2.72:1, then takes over. It also reaches a velocity of 2,400 meters a second. But since it was already moving at the rate of 2,400 meters per second before it was even ignited, its final velocity would be 4,800 meters per second. If there were a third stage with the same mass ratio, its final velocity would be 7,200 meters per second. This would be just about enough to reach orbit.

THE ROCKET AND ITS PAYLOAD

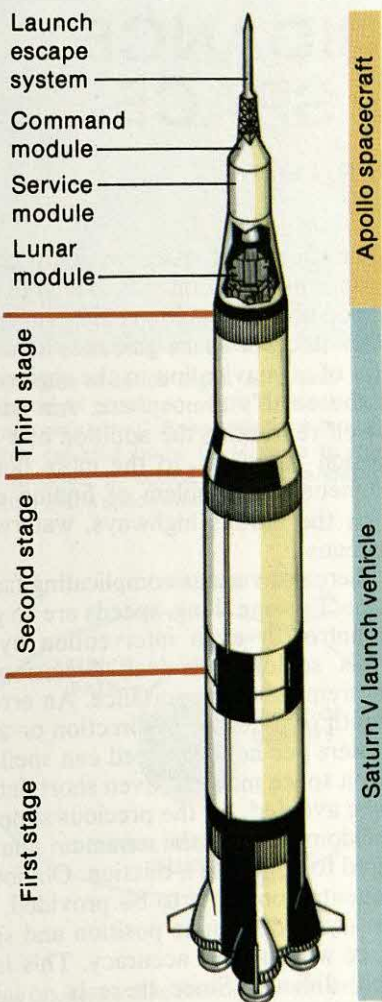
The early pioneers of space flight thought of the rocket as a device that would

The step principle is applied in launching an Apollo spacecraft. The Saturn V launch vehicle consists of three rockets, or stages. The firing of each stage increases the craft's velocity until it is traveling fast enough to enter into orbit around the earth.

someday go into orbit around the earth or perhaps make its way to the moon or to a far-off planet. Nowadays the rocket has come to be considered as the power source that can launch into space a much smaller craft, constituting the payload. The launching rocket is called the launch vehicle, or booster.

Since the liquid-fuel rocket, which is used as a launch vehicle in practically all space flights at the present time, is comparatively thin-skinned, it cannot be fired from a tube or a launching rack. It must be set in place for a launching in a vertical position on a concrete structure called a firing pad. The men who control the launch are stationed in a "blockhouse," which can withstand the shock waves generated by the rocket engine. An official called the range safety officer can destroy in flight any rocket that is not functioning as desired or that has veered badly off course. Naturally this method could not be used if it involved destroying the payload of a manned-rocket flight. Another technique, involving a launch escape tower, is employed.

Considering the fantastic speeds that are eventually attained by the rocket, its flight during the first few seconds after take-off seems slow. When the rocket lifts off the firing pad, it travels less than its own length during the first second. The velocity is about 12 meters per second at the end of the first second; 24 meters per second at the end of the second; 36 meters per second at the end of the third; and so on. The velocities increase rapidly for several reasons. In the first place, the rocket is steadily losing weight as its fuel is being rapidly consumed. Second, the thrust of the rocket becomes stronger as the rocket reaches thinner and thinner layers of the atmosphere, where there is decreasing resistance to the exhaust. (As we have noted, a rocket would work best in an absolute vacuum.) Because of these factors, the rocket



is moving slightly faster than one and one-half kilometers per second one minute after takeoff.

Once the rocket has completed its launching mission, it is detached from the next stage. Naturally the final stage of a rocket that has reached circular velocity and has put the payload into orbit will go into orbit too. It will then—after separation from the payload—form part of the "space junk" that keeps trackers busy. It no longer performs any useful function. It is the payload that carries out the flight mission—exploration, or the establishment and use of orbiting space stations.

GUIDANCE IN SPACE

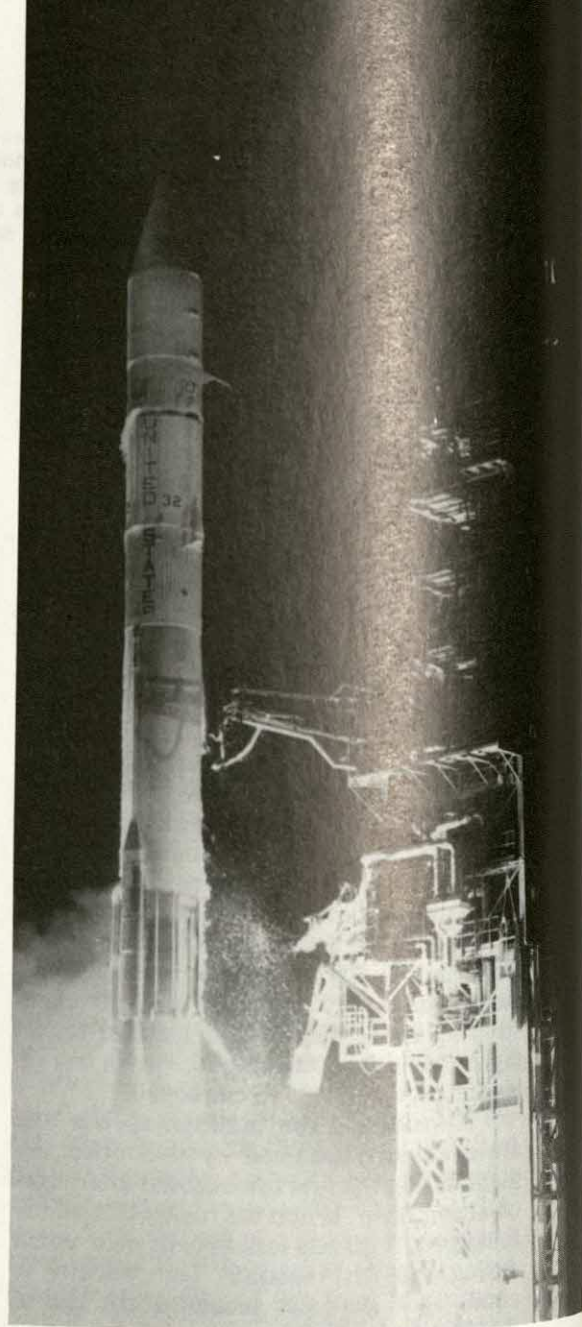
by E. P. Felch

The success of space programs and of long-range missiles depends to a large measure upon ability to guide vehicles in space. We may think of space guidance as an extension of air navigation to the regions beyond the earth's atmosphere. Air navigation itself represents the addition of a third dimension—altitude—to the more familiar two-dimensional problem of finding one's way on the earth's highways, waterways, and oceans.

There are various complicating factors in space. For one thing, speeds are so great that control or even intervention by humans is seldom practical. Precision acquires tremendous importance. An error of one-tenth of a degree in direction or a few kilometers per hour in speed can spell failure for a space mission. Even short detours must be avoided, for the precious supply of fuel seldom exceeds the minimum quantity required to complete a mission. Obviously, if adequate control is to be provided, it is necessary to determine position and speed in space with utmost accuracy. This is extremely difficult. Since there is no air in space, altitude cannot be measured by air pressure and speed cannot be determined by air flow. More complicated procedures, seldom used in conventional means of travel, must be employed.

SPACE-FLIGHT OBJECTIVES

Flights in space have various objectives. They may be summed up as follows: (1) Missile flights to definite target points on the earth (Figure 1, A). (2) Satellites set in orbit around the earth or the sun (Figure 1, B). (3) Space missions to the moon or planets (Figure 1, C). (4) Missiles employed to intercept other missiles (Figure 1, D). (5) Space vehicles rendezvousing with objects already in orbit (Figure 1, E). This may be for inspection purposes or for the



Hughes Aircraft Company

Intelsat IV, a communications satellite, is launched aboard an Atlas Centaur rocket.

assembly of space stations or of particularly large vehicles capable of missions that could not be accomplished by craft launched directly from the earth.

TRAJECTORIES AND ORBITS

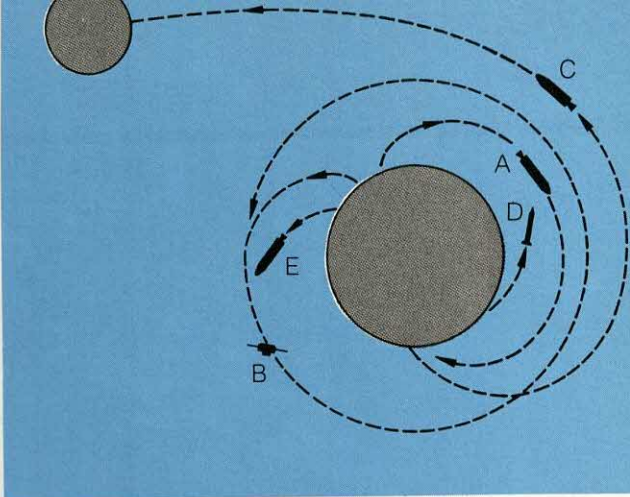
Paths flown in space are usually called *trajectories* if they lead directly from one point to another. The path of a ballistic mis-

sile from launcher to target is a trajectory. So is the path of a space vehicle traveling from the earth to the moon. Flight paths which form closed circuits around the earth or around any other body in space are called *orbits*.

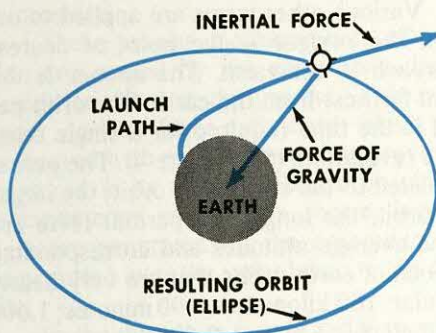
An object coasting in space is said to be in free flight. This means that neither propulsive force, nor friction, nor pressure is acting upon it. Under these circumstances, the flight path of the vehicle is determined by gravitational forces of the earth, moon, sun, or planets acting upon it, and by its own inertia. Inertia is the property which causes a body at rest to remain at rest and one in motion to remain in motion in a straight line at a constant velocity. This tendency toward motion in a straight line, when modified by gravitational forces, results in motion in a closed orbit, forming what is called an ellipse. (Figure 2).

It is quite easy to draw an ellipse. Set two tacks in position about four centimeters apart on a board, and slip a loop made from about 12 centimeters of string over the tacks. Draw the loop taut with the point of a pencil. By moving the pencil clockwise while keeping the string taut, you can draw an ellipse, as shown in Figure 3. Each point where a tack has been placed is called a focus of the ellipse. The major axis is a straight line from one side of the ellipse to the other, going through each focus. The line perpendicular to the major axis and midway between the foci (plural of focus) is called the minor axis. A circle is simply an ellipse in which the two foci fall on the same point and in which the major and minor axes are equal in length. The eccentricity of an ellipse—the measure of how much it departs from the circular form—is related to the difference in length between the major and minor axes. The greater this difference, the greater the eccentricity. In the case of a circle, of course, the eccentricity is zero.

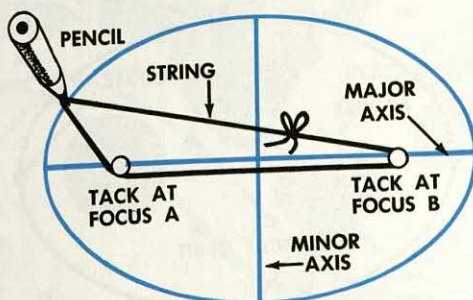
One focus of the orbital ellipse is always located at the effective center of gravitational force. For orbits within a few thousand kilometers from the earth, this may be considered as the center of the earth (Figure 4). The location of the other focus



1. Objectives of space flight: A. Missile flights to definite target points on earth. B. Satellites set in orbit around earth. C. Space missions to the moon or other celestial body. D. Missiles to intercept other missiles. E. Space vehicles rendezvousing with man-made space objects.



2. How a satellite falls into orbit. As it proceeds in its launch path, the satellite is acted on by gravity and enters an elliptical orbit.



3. How to draw an ellipse, as explained in text.



All missile flights and space missions are guided in space from complex control centers such as this one—the Mission Control Center in Houston, Texas, which directed the Apollo flights and landings on the moon.

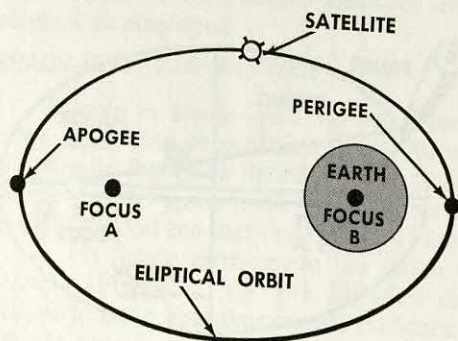
NASA

is determined by the position and motion of the vehicle at the instant when all propulsive thrust comes to an end.

DESCRIBING AN ORBIT

Various other terms are applied to orbits. The *perigee* is the point of nearest approach to the earth. The *apogee* is the point farthest from the earth. The orbit period is the time required for a single complete revolution. (See Figure 4). The period is related to the size of the orbit; the larger the orbit, the longer the period. Here are some average altitudes and corresponding periods of earth orbits that are very nearly circular: 160 kilometers—90 minutes; 1,600 kilometers—2 hours; 9,600 kilometers—5 hours; 36,800 kilometers—24 hours; 384,000 kilometers—28 days. The last pair of numbers relates to the orbit of the moon.

An orbit girdling the earth in a plane parallel to that of the equator is called an equatorial orbit. The so-called polar orbit is



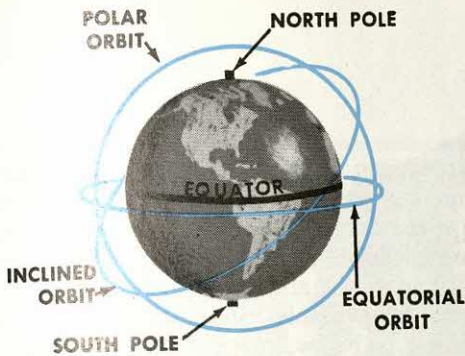
4. The elliptical orbit of a satellite around the earth, with apogee and perigee points indicated.

at right angles to it, passing directly above the North and South poles. The orbits at intermediate angles are termed inclined orbits. (See Figure 5.)

Since orbits are fixed in space, rather than in relation to the earth, the rotation of the earth causes satellites in polar and inclined orbits to pass over different portions of the earth on successive orbits (Figure 6). Since the motion of the earth is precisely known and since that of satellites may be precisely measured after a few orbits, it is possible to predict the paths of orbits over our planet.

Most orbits more than 160 kilometers above the earth are exceedingly stable. However, there are exceptions. For example, satellites with unusually large size-to-weight ratios, such as the *Echo I* 30-meter balloon, are significantly affected by the pressure of the sun's radiation and by collision with the few molecules of air present even at altitudes of 1,600 kilometers. In the course of six months, the apogee of the *Echo I* balloon satellite changed from 1,690 kilometers to 2,170 kilometers, while its perigee changed from 1,530 kilometers to 970 kilometers. Satellites at altitudes of 160 kilometers or less encounter so much of the earth's atmosphere that the orbit altitude decreases rapidly until the satellite either returns to the earth or burns up from air friction in the lower layers of the atmosphere, just as many meteorites do when they pass through the atmosphere.

For many satellite missions, such as for communications and weather, a constant altitude above the earth is desired. In



5. Types of earth orbits: equatorial, polar, and inclined.

these cases, a near-circular orbit is sought. However, even a perfectly circular orbit would not achieve uniform altitude above the earth since the earth is not a perfect sphere. Not only is it somewhat flattened at the poles; it is also slightly pear-shaped.

GUIDANCE SYSTEMS

Thus far, we have been concerned mainly with what is called celestial mechanics—the physical laws that govern the behavior of any body in space. These laws of celestial mechanics represent both tools and limitations for the designer of guidance systems for space vehicles.

The three basic elements common to most guidance systems are (1) sensors; (2) computers; (3) flight controls. Sensors are measuring instruments which determine the actual paths of space vehicles. Electronic computers compare these paths with predetermined paths stored in their electronic memories, and they compute appropriate corrective orders for steering and control-

ling the thrust of engines. Flight controls accept these orders and put them into effect.

Guidance systems for missiles and space vehicles may be divided into two general categories: radio and inertial. We shall discuss some details of each of these systems in the pages that follow.

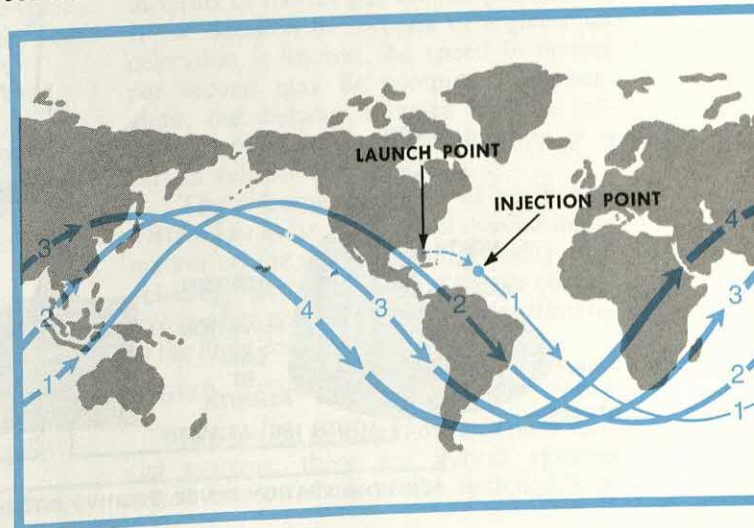
RADIO-COMMAND SYSTEMS

In a radio-command guidance system (Figure 7), the sensor is an amazingly precise radar located on the ground. It measures the position of the space vehicle in terms of angles of elevation and azimuth and also in terms of range. The angular measurements are precise to a hundredth of a degree, while range is measured to within a fraction of a meter, even at hundreds of kilometers.

These precise measurements of position are passed by the radar to a digital computer, also on the ground. This computer derives flight-path data which are compared with programmed data previously stored in its memory. Based on almost instantaneous comparison, corrective orders are generated by the computer and transmitted over the radar beam to the missile.

In the missile, a simple radar receiver receives these control orders, which are in the form of a code. It also triggers a radar transmitter, or beacon, which returns a

6. Successive passes of a satellite in an inclined orbit around the earth. The satellite was launched from a point in Florida. It went into orbit at the injection point indicated. Number 1 represents the first pass; 2, the second; 3, the third; and 4, the fourth.

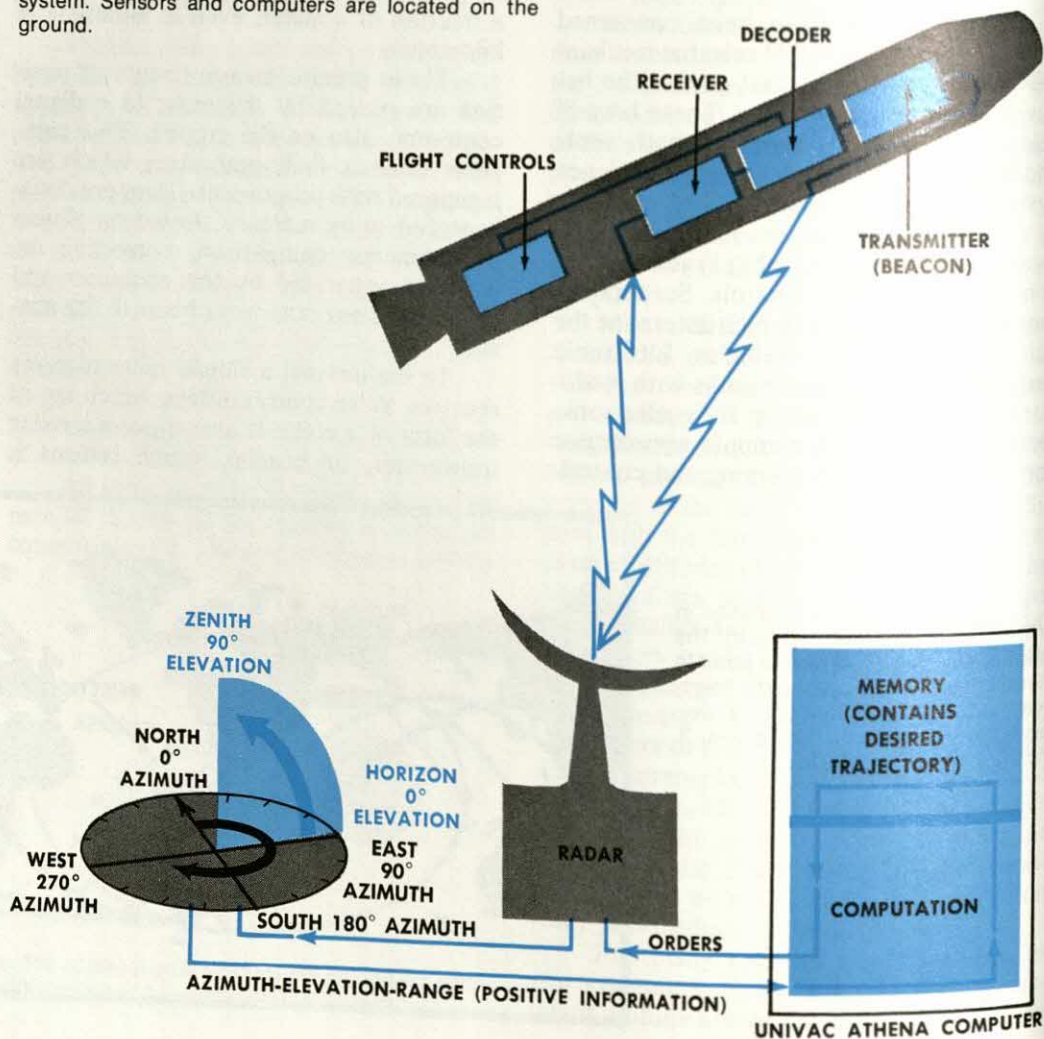


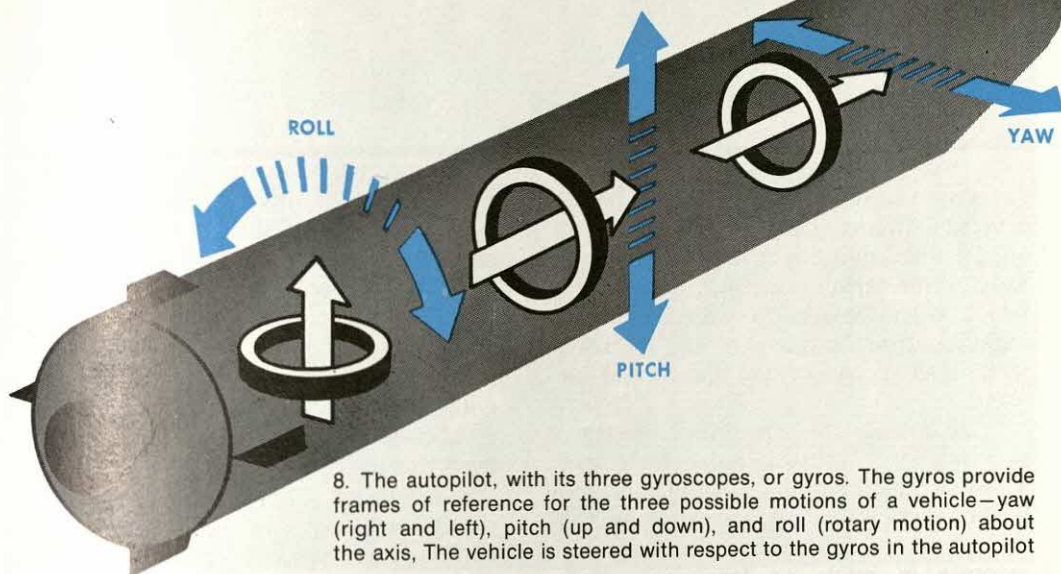
strong signal to the ground to facilitate accurate tracking. A decoder connected to the receiver decodes the control orders and transmits steering and engine commands to the flight controls.

One of the principal advantages of the radio-command guidance system is the light weight of the guidance equipment that is carried aboard the missile. Remarkable reduction in weight is achieved through the use of transistors. We pointed out that the decoder sends steering and engine commands to the flight controls. The steering function of the flight controls is accomplished by a device called an autopilot (Figure 8). This contains three spinning

gyroscopes, or gyros. A gyro, like a child's spinning top, will maintain a fixed position in space even though its support is moved through space. The three gyros of the autopilot provide stable frames of reference for the three possible kinds of motion of the vehicle: (1) yaw, or right-and-left motion; (2) pitch, or up-and-down motion; (3) roll, or rotation about the long axis of the vehicle. All possible maneuvers of the latter are simply combinations of the three motions. In the absence of corrective orders, the autopilot aligns the flight path of the vehicle with the axes of the three gyros. Corrective control orders steer the missile by moving it with respect to the gyros in the autopilot.

7. A simplified diagram of a radio command guidance system. Sensors and computers are located on the ground.





8. The autopilot, with its three gyroscopes, or gyros. The gyros provide frames of reference for the three possible motions of a vehicle—yaw (right and left), pitch (up and down), and roll (rotary motion) about the axis. The vehicle is steered with respect to the gyros in the autopilot.

Steering is usually accomplished either by swiveling the main engine or the smaller auxiliary engines or else by controlling jets of hot or cold gases, aimed in appropriate directions.

The velocity or speed of a vehicle is controlled by shutting off the engines to terminate the thrust at the precise moment when the desired velocity is attained. It is extremely important to determine precise velocity for both ballistic missiles and space vehicles. Intercontinental ballistic missiles must achieve a speed of about 6,100 meters per second, or 21,700 kilometers an hour, to reach a target 8,000 kilometers away. A satellite must reach a speed of about 7,600 meters per second in order to orbit the earth. To travel beyond the grip of the earth's gravity, so-called "escape velocity" of greater than 11,000 meters per second must be attained.

INERTIAL-GUIDANCE SYSTEMS

Inertial-guidance systems employ inertial devices as sensors. In such systems, the sensors, together with the computers and flight controls, are carried aboard the space vehicle.

Gyros are employed as directional references in inertial systems in much the same way as they are in autopilots, as described above. To determine the path of a vehicle in space, velocities must be measured accurately along the directions determined by the gyros.

The inertial devices employed for

measuring velocities are called accelerometers. These actually measure acceleration, or change of velocity, which can be converted into velocity. To understand the principle of operation of an accelerometer, let us consider the behavior of a weight suspended by a string from the ceiling of an airliner (Figure 9). While the airliner is at rest, the weight will hang straight down. When the airliner accelerates on taking off, the weight will swing to the rear. When it decelerates on landing, the weight will swing forward. The distance which it swings is a measure of the acceleration or deceleration, as the case may be. If the length of the string and the distance of the swing of the weight are known, it is possible to calculate the acceleration of the airliner in terms of meters per second per second. If the duration in seconds of a given acceleration is known, the speed in meters per second may be computed. Furthermore, the distance traveled may be calculated quite simply if the duration of a certain velocity is known.

The airborne computer of an inertial system has at its disposal all needed information on the direction and velocity of a vehicle's flight path. It can compute corrective control orders which it passes directly to the flight controls.

HYBRID GUIDANCE SYSTEMS

In addition to pure radio and pure inertial systems, there are hybrid systems which employ some of the techniques of

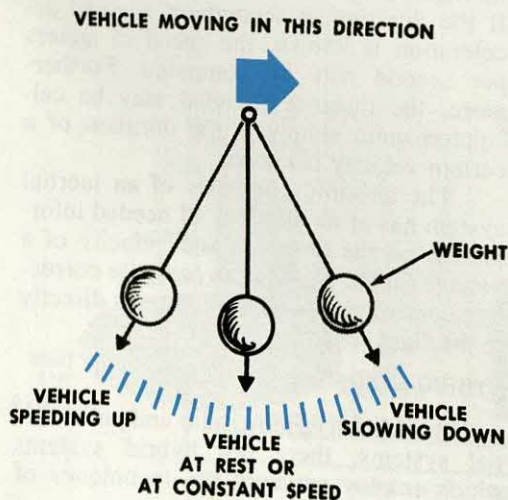
both. Various types of sensors are sometimes used to supplement radar and inertial devices. Among these are "horizon sensors," which provide a reference with respect to the earth's horizon; "sun seekers," which align themselves with the intense radiation from the sun; and "star trackers," which can recognize and aim at star patterns.

Velocities are sometimes measured precisely by Doppler radar techniques. These depend upon the same phenomenon which, applied to sound waves, causes the whistle of an approaching locomotive to increase in pitch, or frequency. Radio waves behave similarly. The change in frequency is an exact measure of the relative velocity between the radio transmitter and the radio receiver. The transmitter is in the space vehicle, the receiver on the ground; or vice versa.

GUIDANCE IN POWERED PHASES OF FLIGHT

Most space vehicles are powered only during a portion of their flight. The rest of the time, they are in free flight, or coasting. While a vehicle may be tracked during any part of its flight, it is obvious that active guidance can be applied only as long as power is available to modify the course.

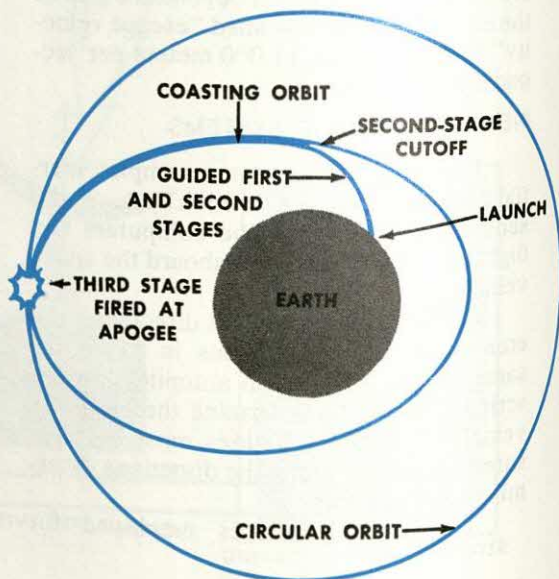
9. The principle of the accelerometer in the inertial guidance system.

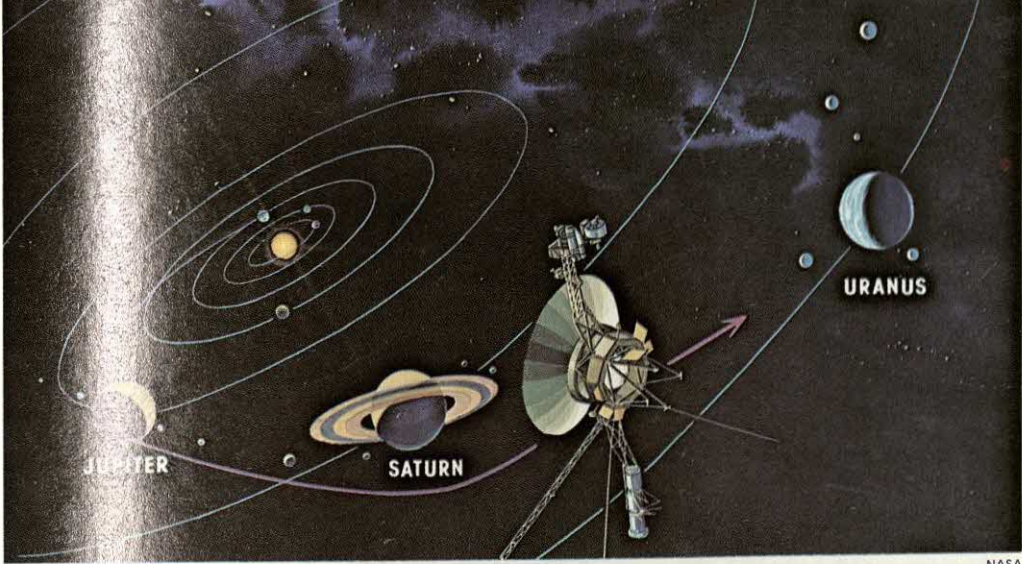


Guidance systems are very versatile. For example, in order to place the 30-meter *Echo I* balloon satellite in a nearly circular orbit about 1,600 kilometers above the earth, it was necessary to resort to what is known as the coasting orbit technique (Figure 10). The third-stage rocket and the balloon were placed in an elliptical orbit with an apogee of about 1,600 kilometers. As apogee was reached, the third stage was fired, and the balloon satellite was placed in the desired orbit. Space missions now often use different types of guidance for the ascent, mid-course, and terminal parts of the flight.

Command guidance systems have already achieved many striking successes. They have been responsible for the successful operation of the many weather, communications, earth observation, and other research satellites launched each year, and for manned and unmanned exploration of space. Perhaps the most outstanding example of the dependability, flexibility, and sophistication of guidance systems was the linkup in space of the U.S. Apollo and the Russian Soyuz spacecraft in July 1975.

10. The *Echo I* satellite was put into a nearly circular orbit around the earth. The three-stage rocket with payload achieved an elliptical orbit with apogee at 1,600 kilometers. When the third stage was fired, the payload entered the desired circular orbit.





NASA

The two Voyager spacecraft flew by Jupiter and Saturn, and then one—Voyager 2—sailed on toward Uranus and a rendezvous with that planet scheduled for 1986.

SPACE SATELLITES AND PROBES

“The Space Age Is Here,” proclaimed the *London Daily Express*. Similar banner headlines appeared on newspapers around the world. Scientists and ham radio operators tried to pick up radio signals. Thousands of people scanned the sky with binoculars, looking for a metal ball 58 centimeters in diameter.

The metal ball was named Sputnik 1. On October 4, 1957, the Soviet Union astounded the world when it launched the 84-kilogram object into an orbit around the earth. Moving at a speed of 28,800 kilometers per hour, Sputnik 1 circled the earth in 1 hour 36.2 minutes. Its two radio transmitters sent continuous signals that were strong enough to be picked up by amateur radio operators.

Sputnik 1 was the first of a great number of unmanned vehicles that have been launched into space. More than 1,000 such craft have been sent aloft since that historic autumn day in 1957. These vehicles have carried out a wide variety of exciting missions and have sent back to earth a great deal of information about our solar system.

PAYLOADS IN SPACE

The object that is carried aloft by a rocket is called the *payload*. The payload may be an unmanned device, bristling with scientific apparatus, or it may be a capsule with one or more astronauts aboard. It is not a rocket, though manned capsules and many unmanned craft are provided with small rocket engines for maneuvering purposes. Whatever the nature of the payload, it is commonly referred to as a *spacecraft* or a *space vehicle*. If it goes into orbit around the earth, it becomes an *artificial satellite*, whether it is manned or unmanned. If the spacecraft is directed far out into space in order to seek information, possibly about a celestial body such as the moon or Venus or Mars, it is called a *space probe*.

The course that the payload is to follow must be carefully mapped out in advance of the launch. This is as true of space vehicles that are to go into orbit around the earth as it is of those that are to be sent far out into space as space probes.

Mapping the course of a satellite. The orbit of any satellite revolving around a heavenly body is normally an ellipse, with the center of that body at one focus of the ellipse. This is true of the earth's orbit around the sun, as well as the moon's orbit around the earth. The orbits of artificial satellites are also generally elliptical, though roughly circular orbits have been achieved in a few cases.

Every orbit has its *apogee*, or point farthest from the earth, and its *perigee*, or point nearest to the earth. These vary widely with different satellites. Sputnik 1 had an apogee of 940 kilometers and a perigee of 234 kilometers. The apogee of the U.S. communications satellite known as Early Bird is 36,373 kilometers, and its perigee is 34,797 kilometers. Sometimes the apogee may be far out in space while the perigee may be comparatively close to the earth. For example, the U.S. craft Explorer 14 had an apogee of 97,904 kilometers and a perigee of 278 kilometers.

Scientists must decide in advance of a flight what sort of orbit is to be achieved and how high it is to be above the earth. The plane of an artificial satellite's orbit must always include the center of the earth, but it may be directed in various ways. The actual orbit, worked out in advance, will depend upon the intended mission of the satellite. The plane of the orbit may take in both poles as well as the center of the earth, in which case the orbit is said to be *polar*. If the plane of the orbit lies along the equator, the orbit is said to be *equatorial*. Intelsat 3-A, a communication satellite, has an equatorial orbit. The orbit may also be inclined by any number of degrees to the plane of the equator. For example, the inclination of Explorer 1 was 33.6°.

Mapping the course to a planet. Several unmanned space probes have passed by or landed on Mercury, Venus, Mars, and Jupiter. In planning the course of such a space probe, it must be kept in mind that the probe must move faster than the earth in its orbit if it is to go away from the sun and must move more slowly than the earth if it is to approach the sun. While the orbit of a neighboring planet could be reached at

any time, it requires careful timing of the shot if the space probe is to reach the other planet's orbit when the planet is also at that point.

It would not do to aim the spacecraft at the particular spot in the heavens—call it point A—where the planet happens to be at the time of the launch. We must remember that it would take a space vehicle months before it would reach any part of the planet's orbit. By the time the craft would have reached point A, the target planet would have reached point B in its orbit, millions of kilometers from point A.

Hence it is necessary to calculate the spacecraft's path so that at a period of some months from the time of launch, the space vehicle and the planet will reach the same point in the heavens. For example, it took the Soviet spacecraft Venera 5 and Venera 6 approximately 130 days to reach the orbit of Venus. Hence their launchings occurred about 130 days before Venus reached the spot where the spacecraft crossed the planet's orbit.

The period during which such a shot can be attempted is called a *launch window*. The launch window for a Venus shot lasts about two weeks; that for a Mars shot, about one month.

Guidance in space. There are many chances for things to go wrong on a space flight. The smallest inaccuracy in calculation or the slightest malfunction of a rocket may cause the space vehicle to veer far off target. This is why each is provided with a guidance system to keep it on course. There are various kinds of systems. In the radio-command guidance system, changes in velocity and direction are broadcast by the spacecraft to a station on the ground. The data are fed into computers which make the calculations required to keep the craft on course. The necessary adjustments are then broadcast to the craft, and control motors on the craft carry out the instructions. The procedure may be directed, at least in part, by ground personnel, or it may be automatic. Some craft are provided with a self-contained inertial-guidance system. This may be actuated by radar beams reflected from the target, such as the moon or

Two and one-half years after Surveyor 3 landed on the moon, an Apollo 12 astronaut visited the probe and removed portions of it for study.



USIS

a planet. Light from nearby stars may be converted into electrical impulses by means of photoelectric cells and may turn on the system.

The end of the flight. The flight of an orbiting artificial satellite of the earth must come to an end eventually, unless its perigee is above the most tenuous part of the atmosphere. Though the atmosphere is rarefied two hundred kilometers from the earth, say, it is not an absolute vacuum. Molecules and other particles exist there. When such particles strike an orbiting spacecraft, the individual collisions will be insignificant, but there will be a cumulative effect. In time the craft will begin to lose velocity. It will not go so high up at apogee, and at its next perigee it will be closer to the earth. The spacecraft will eventually not be able to balance the pull of gravity. It will then penetrate the denser part of the atmosphere.

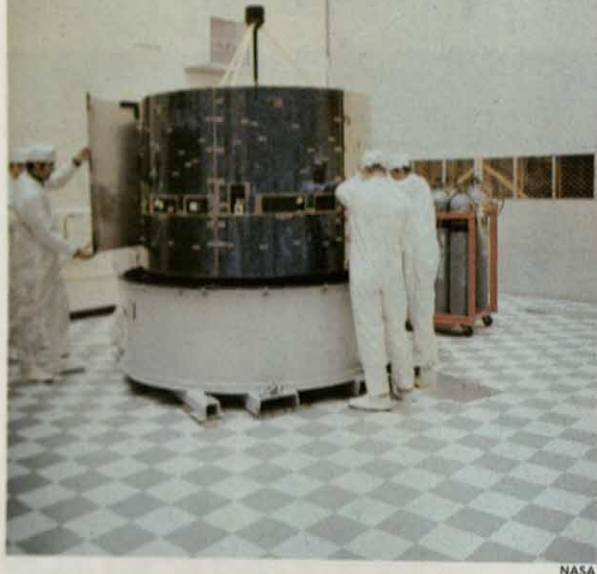
Unless special protective measures or devices, or both, are provided, as in the case of all manned spacecraft and some unmanned craft, a spacecraft reentering the atmosphere is doomed. The same thing will happen to it that happens to meteorites when they penetrate the denser part of the atmosphere. It will be subjected to intense

heat as it collides with the increasingly numerous molecules and ions of the atmosphere, and it will burst into flame. The chances are that it will be utterly consumed. Some fragments, however, may make their way to earth.

A number of capsules, ejected from orbiting unmanned spacecraft, have been parachuted earthward. Some have been recovered by planes in midair. Others have landed on the ocean or on land.

The end of an orbiting space flight may occur in a matter of hours if the perigee is too close to the denser part of the atmosphere. If the perigee is far removed from the earth, the flight may go on for a very long time. It is estimated, for example, that the U.S. satellite Vanguard 1, with an apogee of 4,000 kilometers and a perigee of 640 kilometers, will remain in orbit for a number of centuries. In practically all manned-craft flights, the craft has been deliberately brought out of orbit and returned to earth.

Unmanned craft that escape from the earth's gravitational field may find their end by way of collision. For example, they might crash on the moon. But if they fail to collide with the target planet, they will go into orbit around the sun because every-



The second in a series of Synchronous Meteorological Satellites being prepared for launch. SMS-2 was put into orbit in 1975 to keep a weather watch on the western half of the United States.

thing in the solar system is in the sun's gravitational field. Once they orbit the sun, they will stay in such orbits indefinitely.

VARIETY OF UNMANNED VEHICLES

As we mentioned earlier, more than 1,000 unmanned space vehicles have been launched in the years since Sputnik 1. Each of these craft has had a mission or series of missions. In general, one may say that an unmanned spacecraft is a research device designed to add to man's knowledge of the atmosphere and of outer space or else to solve various practical problems. It may analyze cosmic radiation or cosmic dust. It may count the collisions with meteorites in space. It may be designed to scan earth's cloud formations, so as to map weather changes.

Almost all unmanned spacecraft are designed to transmit information to the earth by means of a telemetering system. In this system, data collected by the scientific instruments on the craft are converted into radio signals, and these are transmitted to ground stations. At these stations, the signals are recorded on spools of magnetic tape. Later they are decoded with the aid of electronic computers. Even photographs can be transmitted in this way.

Unmanned spacecraft require a power source to keep their instruments functioning and to provide the necessary radio transmission. The most commonly used source is a battery of solar cells—waferlike silicon cells that convert sunlight into elec-

tric current. Usually the instruments in a satellite powered by solar cells draw their current from a nickel-cadmium battery, and this battery is recharged by the solar cells. The cells must be exposed to the sun in order to work. Generally, large areas of the satellite are covered with them.

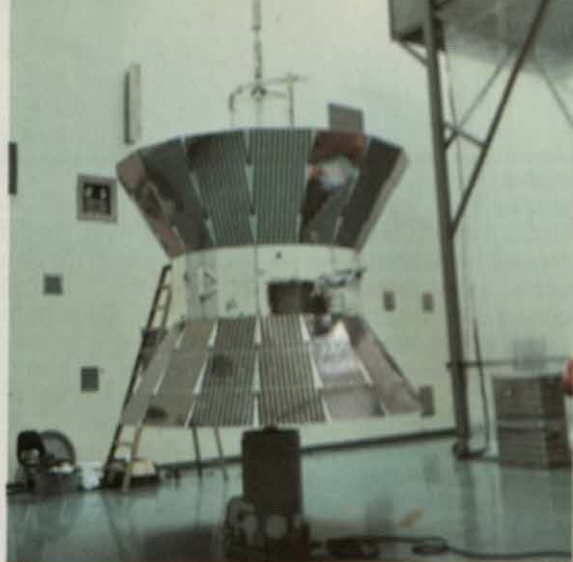
Other power sources, independent of the sun, have been used. For craft designed to remain in orbit for a comparatively short time, ordinary electric batteries may suffice. The SNAP generator, a device that produces electricity directly from atomic energy, has been used with considerable success in the Transit satellites of the U.S. Navy. The device known as the fuel cell has also been used, and is considered most promising. You may have seen the familiar classroom experiment in which water is decomposed into oxygen and hydrogen by an electric current. A fuel cell works in reverse. Liquid hydrogen and liquid oxygen are taken along in separate tanks. They are then recombined chemically, yielding water and producing an electric current.

ADDING TO OUR KNOWLEDGE

Unmanned satellites have vastly increased our knowledge about space and of the bodies in it. Let us briefly consider some of these contributions:

Information about space conditions. A number of satellites have been launched to study conditions in space—radiation, magnetism, dust, meteorites, and so on. The U.S. Pegasus series, designed to report

Helios A, a U.S./German solar probe. It passed within 45 million kilometers of the sun—the closest approach by any such object—and gathered much valuable information.



NASA

punctures by meteorites, has helped engineers design the walls of spacecraft. The Explorer series, the largest group of satellites in the U.S. space program, has provided much data on radiation, magnetic fields, and radio waves in space.

On January 25, 1983, an Infra-Red Astronomy Satellite (IRAS) went into earth orbit. Built around a 57-centimeter telescope and sensors able to detect heat radiation, the U.S.-British-Dutch-operated satellite began an 11-month scan of the entire sky, recording the faint heat of newborn stars, interstellar dust, distant galaxies, and quasars. On April 25 it discovered a comet (IRAS-Araki-Alcock) that approached to within 4.7 million kilometers of earth. An optical space telescope with a 240-centimeter mirror is scheduled for launch early in 1986. It is expected to see objects 50 times as dim and seven times as far out into space as earthbound telescopes see.

Observations of the earth. Variations in the orbits of the Vanguard 1 and 2 satellites (1958–59) showed that the earth is slightly pear-shaped from pole to pole and elliptical around the equator. Since its inception in 1964, the OGO (Orbiting Geophysical Observatory) series has provided information on the sun's influence on the earth and space.

One of the most exciting discoveries made possible by unmanned satellites was the discovery of a vast zone of radiation encircling the earth above the equatorial

regions. Within the zone, charged particles derived mostly from the sun are trapped by the magnetic field of the earth. Satellites 1, 2, and 12 of the Explorer series provided the data that led to the discovery and analysis of the earth's radiation belt.

The natural resources of the earth are being surveyed by ERTS—the Earth Resources Technology Satellite. It takes pictures of earth and sea in near-infrared radiation and in the red and green light bands of the spectrum.

Another important development in earth observation occurred on July 23, 1972, when the first Landsat satellite was launched. By 1983, three additional Landsats had been placed into orbit. Landsats make complete earth orbits every 103 minutes, recording images electronically. These electronic signals are returned to earth, where they are converted by computer into visual images. Landsat is used to map the earth's surface features; assist in locating mineral resources, including oil reserves; and chart geological faults in the hope of developing an early-warning earthquake system.

In 1979 a magnetic field satellite, Magsat, was launched to survey the earth's magnetic field. The satellite plunged into the atmosphere in June 1980 after completing its mission.

Observations of the sun. Intensive observations of the sun have been carried out with the aid of the Orbiting Solar Observatory (OSO), satellites of the Explorer and

Pioneer series, and other spacecraft, including two West German solar probes, Helios 1 and 2, launched in 1974 and 1976. Studies have been made of the solar particles, solar flares, ultraviolet rays, the corona, and the solar wind.

Later studies included the launching in 1980 of Solar Max (Solar Maximum Mission), a 575-kilometer-high earth-orbit observatory of the sun. Solar Max is designed to study the physics of solar flares and their effects on earth. In 1981 a Mesosphere Explorer satellite was launched to study the sun's effect on the formation and destruction of ozone in earth's upper atmosphere.

Weather satellites. These devices offer the advantage of showing cloud formations over large areas of the earth's surface by means of pictures taken with television cameras and relayed by telemetering to the earth. The United States has launched more than 20 weather satellites since the first, Tiros 1, was sent aloft on April 1, 1960. Ten Tiros satellites were launched, and their cameras provided the first large-scale weather photographs of earth.

An improved type of weather satellite, Nimbus 1, was launched on August 28, 1964. The Nimbus craft have tested a wide variety of experimental weather-monitoring devices. Nimbus 6, which was launched June 12, 1975, measures radiation in the earth's atmosphere, data important to determining climate changes. Meanwhile, the U.S. National Oceanic and Atmospheric Administration has launched Synchronous Meteorological Satellites and GOES-1 as part of its weather-observing system. The U.S.S.R. also has launched a series of weather satellites.

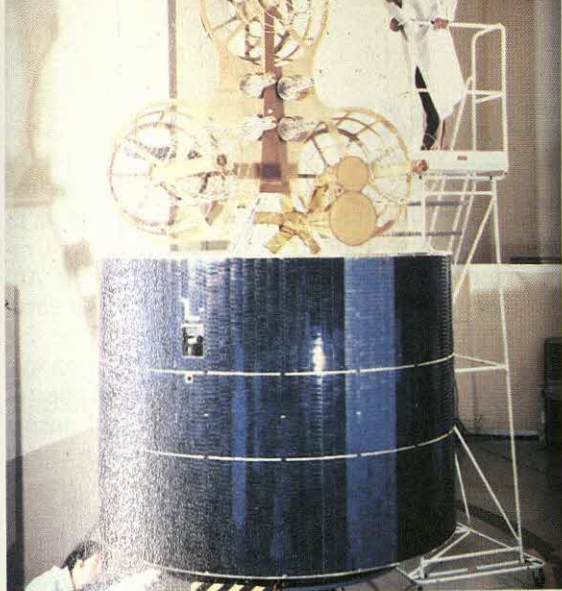
Communications satellites. Earth satellites now provide radio and television service to much of the world. The passive satellites simply bounce signals from one earth station to another without amplification, as the Echo balloons have done. Active satellites amplify the relayed signals. These include the United States' Relay, Telstar, Syncom, and Early Bird; the international Intelsat craft; Canada's Anik series; and the U.S.S.R.'s Molniya satellites. Communications satellites have also

been used to relay educational and health information to isolated villages. In 1976 the joint Canadian-U.S. Communications Technology Satellite, the most powerful communications satellite ever built, was launched; and a Marisat satellite providing ship-to-shore telecommunications for the international shipping industry and the U.S. Navy began operation.

Navigation satellites. The Transit satellites, orbited by the U.S. Navy, are intended to serve as useful aids to navigation when a surface vessel finds itself in an area of bad weather or bad visibility. The fog-bound navigator contacts an orbiting Transit satellite by radio. The satellite then responds, also by radio, with a statement of its position in its orbit. Once the navigator has this information, all he needs, in order to determine his own position, in addition to the customary tools of his trade, is a list of the satellite's orbits.

Biological satellites. Late in the year 1966, the United States began launching a series of Biosatellites. This program is designed to test the reactions of various organisms to space travel. The influence of weightlessness, radiation, and removal from a day-night cycle are being studied. The organisms on these missions have included a wide variety of plants and animals, ranging from bacteria and frog eggs to wheat seedlings and a 6.4-kilogram monkey.

Military satellites. A certain number of unmanned artificial satellites now orbiting the earth are conveying information of military importance to the two nations that up to now have almost monopolized space exploration: the United States and the Soviet Union. Thus, the Midas satellites of the United States can spot the launching of missiles through the use of infrared sensors. The U.S. Samos have taken highly detailed views of Soviet installations. Similarly, the U.S.S.R., with vehicles such as the Cosmos series, photographs U.S. airfields, munitions factories, missile-launching sites, and so on. In 1968 the United States launched a series of highly sophisticated IS (integrated systems) satellites that carry instruments capable of detecting infrared radiation.



Hughes Aircraft Company

Marisat is being used by the international shipping industry and the U.S. Navy as a telecommunications satellite, providing an instant link from shore stations to ships at sea.

The so-called "spy-in-the-sky" satellites received unexpected attention in September 1983, following the aerial destruction of Korean Air Line's flight 007 off the Sea of Japan by Soviet jet interceptors. Detailed knowledge of Soviet activities as well as an accurate recording of Soviet voice transmissions led many observers to believe that U.S. spy satellites had achieved a high level of sophistication.

TO THE MOON

Among the most exciting features of the space programs have been the space probes, which have provided close-ups of celestial bodies.

The Luna program, begun by the U.S.S.R. in 1959, has included a variety of moon probes. Some craft crashed into the lunar surface. Others flew past the moon or went into lunar orbit. For example, Luna 3, launched in October 1959, was the first craft to go completely around the moon. It photographed the side not visible from the earth and transmitted the first photos of this area. Beginning with Luna 9, which was launched in January 1966, several Luna probes have soft-landed on the moon. These have transmitted pictures, density data, and other information. In 1970 and 1972, respectively, Lunas 16 and 20 scooped up moon rock and dust samples, to send them to earth by rocket. In 1970 and 1973, Lunas 17 and 21 landed Lunokhods 1 and 2—unmanned wheeled vehicles

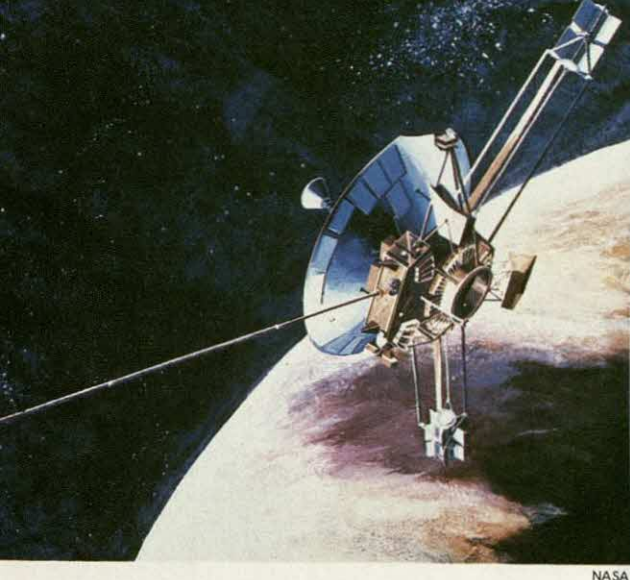
that were driven by remote control from earth. The Soviet Zond program has also included lunar missions.

The United States has also sent a variety of probes to the moon. Rangers 7, 8, and 9 (1964–65) took more than 17,000 pictures of the moon's surface before they crashed upon it. Unmanned satellites, known as Lunar Orbiters, orbited the moon, photographing its surface, and a series of Surveyors soft-landed on the moon, testing its surface. Among other U.S. lunar probes is an Explorer 49 radio-astronomy satellite launched in 1973 to circle the moon. The most exciting program involving the moon has, however, been the U.S. Apollo project, which landed men on the moon. To learn more about this project, see the articles "Manned Space Flight" and "Exploration of the Moon" in *The New Book of Popular Science*.

TO THE PLANETS

The nearest planets, Venus and Mars, are millions of kilometers away, the outer planets hundreds and thousands of millions of kilometers from earth. Probes have, however, reached some of the planets and are approaching others.

The United States Mariner program has studied three planets: Mercury, Venus, and Mars. Mariner 2, launched in August 1962, passed within 35,200 kilometers of Venus and transmitted important information about the planet to the earth.



As Pioneer 10 flew past Jupiter, it sent back pictures that gave us our first close-up of the planet. The probe is now leaving the solar system.

Among other things, it reported that the surface temperature of Venus is close to 430° Celsius, and that the planet almost lacks a magnetic field. Mariner 5 passed within 4,000 kilometers of Venus in October 1967 and sent to earth information about that planet.

The Soviet Union launched a series of probes—the Veneras—to land capsules on the surface of Venus. Veneras 4 to 8 succeeded in doing so, from 1967 through 1972. All except Venera 4 transmitted data to earth, revealing that the planet's surface is very hot and has an atmospheric pressure 100 times that of the earth's surface.

Mariner 10, launched in 1973, provided new data on Venus and Mercury. After taking the first clear pictures of the swirling atmosphere of Venus, it passed Mercury three times—in March and September 1974 and again on March 16, 1975, when it came within about 320 kilometers of the surface of the planet. The thousands of photographs taken and other data collected provided man's first detailed look at Mercury.

A Pioneer-Venus mission in late 1978 probed the dense atmosphere, high temperatures, and sulfuric mists of earth's "sister" planet. Mariner 4, launched in November 1964, passed within 10,000 kilometers of Mars. It revealed an ionosphere, measured radiation, and transmitted pictures of the Martian surface.

Voyager 2, launched in 1979, is one of several major U.S. space probes to the planet Jupiter. Voyager 2 is expected to pass close enough to the outer planet Uranus in January 1986 and to Neptune in August 1989 to record and transmit to earth extremely valuable scientific data.

At the same time, the Soviet Union was having only partial success with probes to Mars. Several of its probes proved unsuccessful, but Mars 5 took many pictures and Mars 6 sent back data until it landed.

Early in 1969 Mariners 6 and 7 were launched toward Mars and returned photos of its surface. Late in 1971 Mariner 9 reached the vicinity of Mars while that planet was being swept by a huge dust storm. After the dust settled, the Mariner transmitted many pictures of Mars, revealing a world with many complex features.

Two Viking probes landed on Mars in the summer of 1976. The craft transmitted many photos of the planet's surface as well as much important data regarding the chemical composition of the surface. The probes also scooped up soil samples and analyzed them for any indications of life.

Exploration of the outer planets began in 1972 and 1973, when the United States sent Pioneers 10 and 11 toward Jupiter. The probes also supplied important information about asteroids, the solar wind, cosmic rays, and other radiation they met along the way. Pioneer 10 passed near Jupiter late in 1973 and took excellent pictures. The probe revealed that the giant planet has strong radiation belts and a large magnetic field, consists mostly of liquid hydrogen, and is hot inside.

On June 13, 1983, Pioneer 10 crossed the orbit of Neptune and became the first probe to leave the solar system. During its travels to deep space, it made the first accurate measurements of Jupiter's moons, and recorded the first close-up pictures of its Great Red Spot.

While Pioneer 10 flees the solar system, more than 50 other space probes are trapped within the solar system. These include a variety of spacecraft launched by both the United States and the Soviet Union.

COMMUNICATIONS SATELLITES

by J. Kelly Beatty

Imagine that you have just picked up the telephone to call a friend who lives in England. After dialing the number, you wait a few seconds and perhaps hear a distant click or two before the phone starts ringing. Then, after talking for a few minutes, you exchange good-byes and hang up.

If you ever do make such a call, your conversation will probably be sent across the ocean by way of a spacecraft circling high above the earth. Such spacecraft are called communications satellites, or *satcoms*. They are being used increasingly to handle long-distance telephone, television, and other transmissions around the world. For example, they are used for about two-thirds of all transatlantic telephone calls. Of the payloads placed in orbit by the U.S. National Aeronautics and Space Administration (NASA) in 1982-83, more than half of them—21 in all—were communications satellites.

This trend is still growing rapidly, and with good reason. Communications satellites can be used when other forms of communication are either too expensive or impossible. For example, in 1983 seven undersea cables linked North America and Europe. The last one cost more than \$175 million and accommodates 4,200 telephone circuits. In comparison, a modern communications satellite costs roughly \$80 million (including the payment for launching it into orbit), and it can often handle more than three times as many calls. These orbiting switchboards have helped to reduce the cost of long-distance telephone calls dramatically since the 1960's.

EARLY PASSIVE SATCOMS

The first true communications satellite, Echo 1, was launched August 12, 1960, less than three years after the very first satellite, Sputnik 1, was put into orbit. The Echo satellite was actually a plastic balloon 30 meters in diameter, covered with a thin

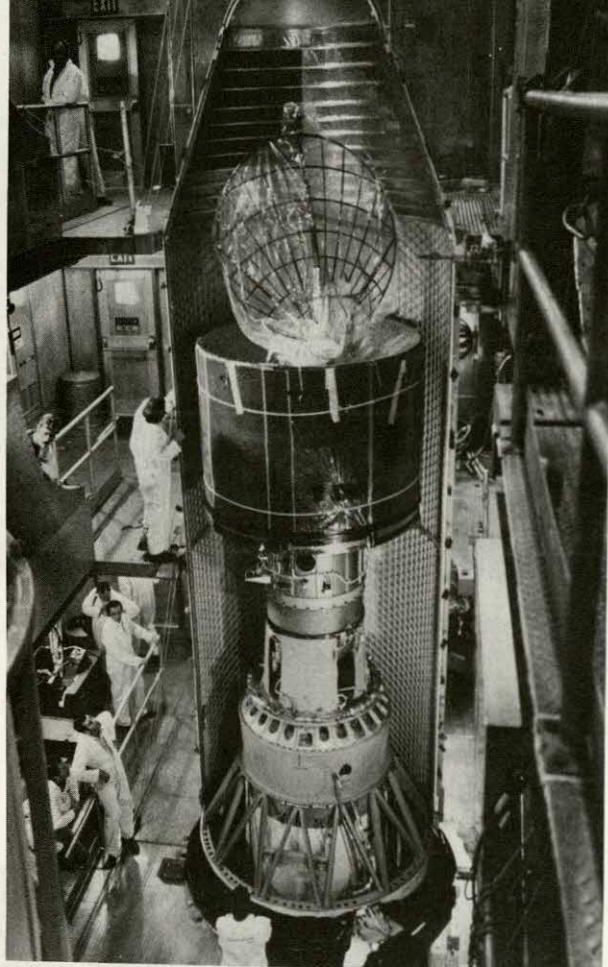
coating of aluminum to reflect radio transmissions. A second, larger Echo was orbited in 1964. Communications satellites such as Echo are called *passive*. They do not amplify radio signals but merely reflect them. They are not really practical, because of their large size and the powerful equipment required to send and receive the messages.

However, the development of passive reflectors did not end with Echo. In one program, called Project West Ford, a thin ring of fine wire needles was spread around

Composite photo shows placement of NASA's Tracking and Data Relay Satellite in the payload bay of the Challenger space shuttle launched in April 1983.

NASA





Hughes Aircraft Company, photo by NASA

Anik is the name of Canada's domestic communications satellite system. Here an Anik craft is being readied for launch at NASA's Kennedy Space Center in Florida.

the earth in 1963. The wires were "tuned" to reflect signals at a certain frequency. However, the program was abandoned because the ring of wires threatened to disturb astronomical observations.

Another novel idea was tried in 1966. An Echo-type balloon was covered with a precisely spaced wire mesh. The balloon decomposed after a few hours in orbit, leaving a hollow metallic sphere that was five times more reflective than Echo, at radio frequencies. It was also much less affected by the sun's radiation and was slowed down less by the atmosphere.

ACTIVE SATCOMS

The commercial value of satcoms was tested in 1962, when Telstar 1 was placed

in orbit for Bell Telephone. Telstar was an *active* satellite, amplifying and retransmitting as many as 60 two-way telephone conversations at one time. Ground stations were established in the United States, England, and France. A few months later, Relay 1 was launched for the RCA Corporation of America, adding Italy and Brazil to the growing list of countries that are now receiving broadcasts from satellites in outer space.

Syncom 2, launched in 1963, became the first communications satellite to achieve a *synchronous orbit*. In a synchronous, or same time, orbit a satellite completes one circle of the earth in the same time as the earth completes one rotation on its axis. An altitude of 35,680 kilometers provides a synchronous orbit. If the spacecraft is traveling in the plane of the equator, it remains over the same position on the earth's surface at all times and is in constant view of almost half the planet. Modern communications satellites are placed in orbits of this type.

From time to time, NASA has participated in programs to test experimental communications systems. The first of NASA's Applications Technology Satellites, ATS-1, was launched in 1967 to test a variety of instruments, including a 600-channeled repeater that could relay color television programming, as well as radio transmissions between aircraft and ground stations.

Many countries have developed or purchased their own communications satellites, with numerous others planning to do likewise in the near future. The Soviet Union began building its orbiting system in 1965 and now has three series of communications satellites (called Molniya, Stationar, and Gorizont) circling the earth. These bring telephone and television service to remote areas of the country, as well as to parts of Eastern Europe. The Canadians and Japanese have also established sophisticated satellite systems for domestic transmissions.

Dozens of other special-purpose satellites have been launched. Today it is not unusual for a country like India or Mexico to purchase a complete system for orbital

communications, which includes the satellite itself, the use of a launch vehicle like the space shuttle, and equipment on the ground (such as receiving antennas) to operate the satellite. More and more, large companies with extensive telephone and computer needs are turning to communications satellites as a practical and economical alternative to conventional transmission methods. Amateur radio operators, or "hams," have built several Orbital Satellites Carrying Amateur Radios (OSCARs) that take advantage of leftover payload space on other missions to get a free ride into orbit. Other communications satellites listen for distress signals from ships and aircraft, then relay the victims' location to rescue teams.

INTELSAT

By far, the most extensive satellite system is that of the International Telecommunications Satellite Consortium, or Intelsat. Intelsat is made up of 109 nations and operates some 500 ground stations worldwide. Beginning in 1965 with the Early Bird satellite, it has steadily expanded its capability and now controls 16 satellites—with more planned in the years ahead. The latest in its spacecraft series, called Intelsat V, can handle 12,000 telephone circuits and two color television transmissions simultaneously. Even more powerful satellites, the Intelsat VI series, will be able to relay more than 30,000 calls when introduced in 1986.

The American member of Intelsat, the Communications Satellite Corporation (Comsat), was established by the U.S. Congress in 1962. Comsat and its subsidiaries provide a wide variety of public and private communications services throughout the United States.

SATCOM OPERATION

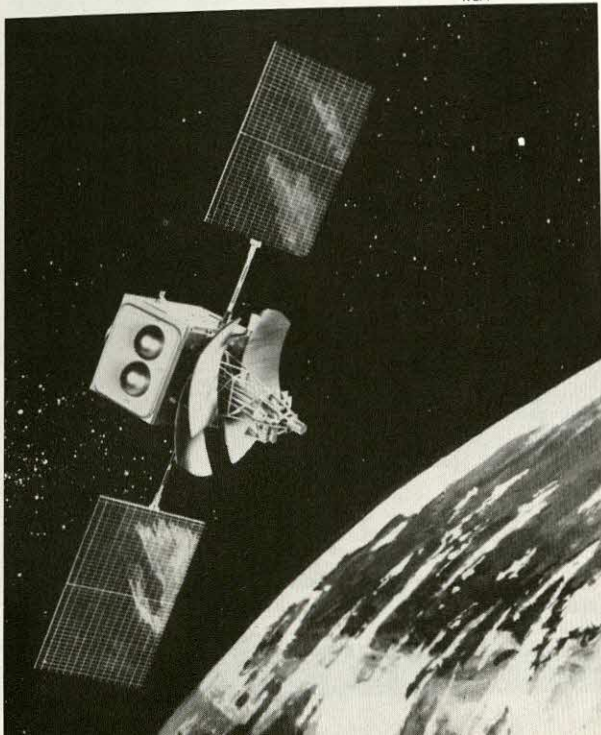
A "typical" communications satellite is usually launched by NASA, though the Europeans, Japanese, and other countries can now provide their own rockets. Once in its prescribed position in synchronous orbit over the earth, the spacecraft is ready to go to work. Ground stations may belong to a country's government, a consortium

like Intelsat, or even private individuals. Messages are usually beamed up to the satellite at one frequency and received at another by using an amplifying device called a *transponder*. This allows a single ground station to send and receive signals at the same time.

Depending on the satellite, its capability can be used for telephone service, encoded data, or television transmissions. The various channels are usually leased to users in the form of *half-circuits*—two-way connections between the satellite and a ground station. Thus, a pair of half-circuits are required to complete an overseas call. These connections are normally operated full time. Other channels are used for occasional television transmissions such as news stories and sports events. Although not in service full time, many TV programs schedule satellite time on a regular basis. A number of countries also rent the use of a transponder for domestic communications or other purposes. Finally, a portion of the circuits are held in reserve for emergencies, increased demand within the system, and other special needs.

An artist's conception of Satcom I in orbit around the earth. This satellite is equipped with transponders, or frequency changers.

RCA



WHERE DO WE GO FROM HERE?

As the need for global communications increases, new communications satellites will be developed to handle the load. High-speed computers, for instance, can quickly relay vast amounts of data via orbiting spacecraft. Cable-television networks use satellites to bring a wide variety of programming to their subscribers. Soon a remarkable new breed of satellites will begin

operation. Their powerful transmitters will beam signals to antennas less than one meter in diameter, making it possible for individual families to "dial in" the satellite channel of their choice.

Before 1956, ships and airplanes were the only dependable means of transoceanic communication. The development of communications satellites has changed all that. Today they make calling a friend in England just as easy as calling one next door.

COMMUNICATIONS SATELLITES *

| Name of satellite or series | Operator | Number | First launch | Mission description |
|-----------------------------|----------------|--------|--------------|--|
| Score | NASA | 1 | 12/18/58 | Broadcast prerecorded voice messages |
| Echo | NASA | 2 | 8/12/60 | Passive reflector |
| Oscar | Amateurs | 7 | 12/12/61 | Used to relay signals from widely separated amateur radio stations |
| Telstar | AT&T | 2 | 7/10/62 | Experimental commercial satellite; 60 voice circuits, or 1 TV channel |
| Relay | RCA | 2 | 12/13/62 | Ground stations in U.S., England, France, Italy, and Brazil; 2 transponders |
| Syncom | NASA | 3 | 2/14/63 | Proved usefulness of synchronous orbits for communications purposes |
| Intelsat I | Intelsat | 1 | 4/6/65 | Service between North America and Europe; 240 voice circuits, or 1 TV |
| Intelsat II | Intelsat | 4 | 10/26/66 | Introduced multipoint connections within area of coverage |
| Intelsat III | Intelsat | 8 | 9/18/68 | Simultaneous telephone and TV service; 1,500 voice circuits or 4 TV channels |
| Intelsat IV | Intelsat | 8 | 3/26/71 | About 3,500 voice circuits and 2 TV channels; 12 transponders |
| Westar | Western Union | 2 | 4/13/74 | Domestic service; 12 transponders |
| ATS-6 | NASA | — | 5/30/74 | Experimental satellite for medical and educational tests in U.S. and India |
| RCA Satcom | RCA | 2 | 12/12/75 | Three-axis stabilized, using sun and star sensors; 24 transponders |
| Statsionar | USSR | 20? | 12/22/75 | Three-axis stabilized; advanced domestic service through 1980's |
| CTS | Canada-NASA | 1 | 1/17/76 | Cooperative experimental satellite; high-power transmitter for broadcast to portable antennas |
| Intelsat IVA | Intelsat | 6 | 2/1/76 | About 6,300 voice and 2 TV channels; band separator for frequency reuse |
| Marisat | Comsat | 2 | 2/19/76 | Maritime satellite for ship-to-shore communications |
| Comstar | Comsat General | 3 | 5/13/76 | Leased full time to AT&T for domestic service; 24 transponders |
| Intelsat V | Intelsat | 9 | 12/6/80 | About 12,000 voice circuits and two TV channels; spot antennas allow reuse of frequencies up to four times |
| SBS-3 | SBS | 3 | 11/11/82 | One of two satcoms launched during first commercial space shuttle mission (STS-5). Third in a series owned by Satellite Business Systems, a private consortium |
| BS-2a | Telesat-Japan | 2 | 2/84 | First operational direct-broadcast satellite to beam TV programs to Japanese homes having one-meter-diameter antennas |

* Military satellites not included

MANNED SPACE FLIGHTS

by Willy Ley

"Tranquility Base here. The *Eagle* has landed."

And so, on July 20, 1969, at 4:17:40 P.M. Eastern daylight time, two men landed on the moon. As Neil A. Armstrong, commander of the mission, relayed the anxiously awaited news across 385,000 kilometers of space, mankind realized an ancient dream—a dream of flying into space and visiting another celestial body. But though the dream was an ancient one, the scientific efforts to achieve the dream began just fifty years before the flight of Apollo 11.

A COMPARATIVELY LATE DEVELOPMENT

At the time that Robert H. Goddard began the period of serious research in space flight with his *Method of Reaching Extreme Altitudes* (1919), radio transmitters and receivers were large and bulky. Telemetry was in the experimental stage; so was television. Since it would then have been impossible to transmit scientific data and pictures from an unmanned spacecraft to the earth, it was assumed that all flights into space would be manned flights.

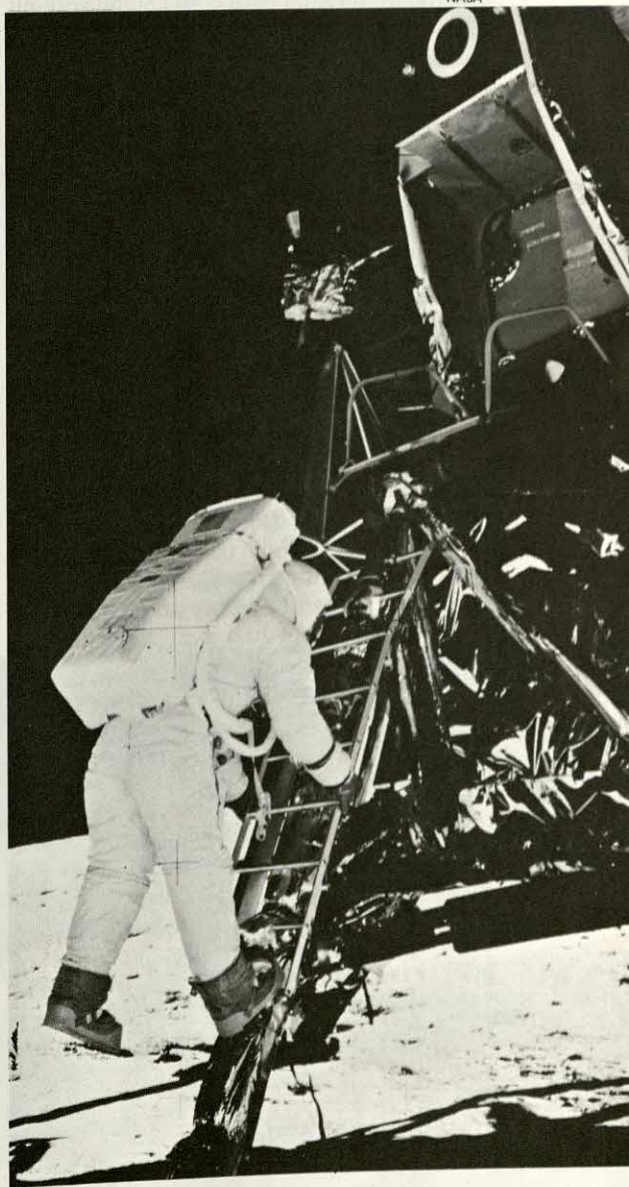
Actually, however, manned flights came rather late in the day. It was not until April 21, 1961, that the first manned flight into space took place, with the Soviet space vehicle Vostok 1 completing one orbit around the earth. By that time, a good many unmanned craft had already been launched.

There were several reasons for this state of affairs. Spectacular advances had been made in radio, telemetry, and television. They had made it possible to guide unmanned spacecraft in ways that had not been dreamed of earlier and also to transmit all kinds of scientific data and even pictures to ground stations. Manned flight, however, could not be achieved until cer-

tain vital problems had been solved, at least in part. Among these problems were excessive acceleration, weightlessness, radiation in space, the dangers of collisions with meteorites, and reentry into the earth's atmosphere.

Excessive acceleration and weightlessness. Gravity constantly pulls us toward the earth's center with a force equivalent to an acceleration of 9.8 meters per second

NASA



Apollo 11 astronaut Edwin Aldrin, Jr., steps off the ladder of the lunar module onto the surface of the moon. He and Neil Armstrong were the first men to walk on the moon.

per second. This rate is called 1 g ("g" standing for "gravity"). Here on the earth, we are no more aware of gravity than we are of the fact that the atmosphere weighs down upon us with a pressure of about 1 kilogram per square centimeter. But if we are subjected to an acceleration of 2 g's, we have a very curious sensation—as if we had suddenly become twice as heavy. This effect is heightened as the acceleration continues.

It was calculated that anyone venturing aloft in a spaceship that was to go into orbit would be subjected to as much as 7 or 8 g's. Experience in flying seemed to indicate that the top limit that a pilot could stand without "blacking out" (becoming unconscious) was 4 g's. How then could future astronauts withstand even greater acceleration—as much as 7 or 8 g's?

To test the ability of humans to tolerate a comparatively high rate of acceleration, centrifugal force was substituted for acceleration. A gondola was suspended from the end of a large arm, which in turn was attached to a rotary motor. A volunteer entered the gondola. As the motor was set going, the gondola sped round and round with increasing velocity. Volunteers who remained in a sitting position underwent alarming effects. The blood circulation slackened, and when it failed to reach certain vital areas of the brain, the volunteer would black out. But if he lay on his back with his head slightly raised and the knees slightly bent, the physical effects of acceleration were not nearly so pronounced. It became evident that a more-or-less supine, or on-the-back, position would have to be adopted during moments of extreme acceleration.

What of the effects of weightlessness on humans? By way of introduction, it should be pointed out that when we say that both objects and persons are weightless in space, we do not mean that the earth's gravity has no effect on them beyond the earth's atmosphere. As a matter of fact, at a height of even 320 kilometers or so, the force of gravity is about 90 per cent as strong as at sea level. It becomes half as strong only when we have risen to a height of more than

2,600 kilometers above the earth's surface. Even far beyond that point, the force of gravity makes itself felt. After all, the moon goes around the earth only because a mutual gravitational attraction holds it at an average distance of about 385,000 kilometers. But there is a great difference between being under the gravitational influence of a body and feeling weight. Weight is felt only if the pull is resisted. If a body follows the pull of gravity freely without resisting it in any way, it is weightless. An artificial satellite that is in orbit around the earth does follow the gravitational pull. If it did not, it would go off into deep space along a straight line. At the same time, it does not resist this pull, since its velocity in orbit counterbalances exactly the effect of the earth's gravity. Therefore the satellite is weightless.

Early investigators caused a state of weightlessness to be produced by means of an airplane maneuver. The pilot first made a shallow power dive, then pointed the nose of the plane upward while throttling back his engine. The plane went through a curve, moving upward at first, then leveling out, and finally pointing downward. Since the shape of the curve was quite similar to that of a parabola, the flight was called parabolic. From the instant of engine-power reduction until full engine power was restored, the plane and everything it contained were weightless. The periods of weightlessness ranged from 25 seconds to about a minute.

During this period, human "guinea pigs" in an experimental chamber in the plane floated around aimlessly. So did all loose objects. Weightlessness also produced physiological symptoms in men. It brought about a sense of disorientation (losing one's bearings), dizziness, and erratic control of one's movements. These effects varied widely in different individuals. They seemed to be slight in some, overpowering in others.

Clearly a period of training would be required to condition future astronauts to endure extreme acceleration and weightlessness. Animals were put through such a program of training, and then were sent aloft in capsules that went into orbit. They

returned safely to earth, generally apparently none the worse for their experience. It was concluded that well-trained astronauts should also be able to withstand increased acceleration and weightlessness.

The effects of radiation. It was well known to investigators that spacecraft in outer space would be subjected to various kinds of radiation. Ultraviolet rays from the sun would strike the craft with undiminished force, since they would not have been absorbed by the earth's atmosphere. Solar flares, associated with sunspots, would send out X rays and electrons into space. X rays would be produced as electrons present in space struck the metal of the spacecraft's walls. (This is known as impact radiation. It would correspond to the impact of electrons striking the "target" in an X-ray tube.) Cosmic rays would also be encountered in space. The discovery of the earth's radiation belt in 1958 laid bare another source of potential danger.

It was found, however, partly through investigations carried out with unmanned satellites, that radiation did not offer an insuperable obstacle to space flight. Ultraviolet rays could be stopped by even very thin sheets of metal. So could X rays coming from the sun, since they were not very powerful. Under normal conditions, the walls of the spacecraft would be thick

enough to shield its occupants against impact radiation. It would not be practicable to provide shielding against cosmic rays; but normally there would not be enough of this type of radiation to constitute a hazard. The earth's radiation belt could be avoided with proper planning.

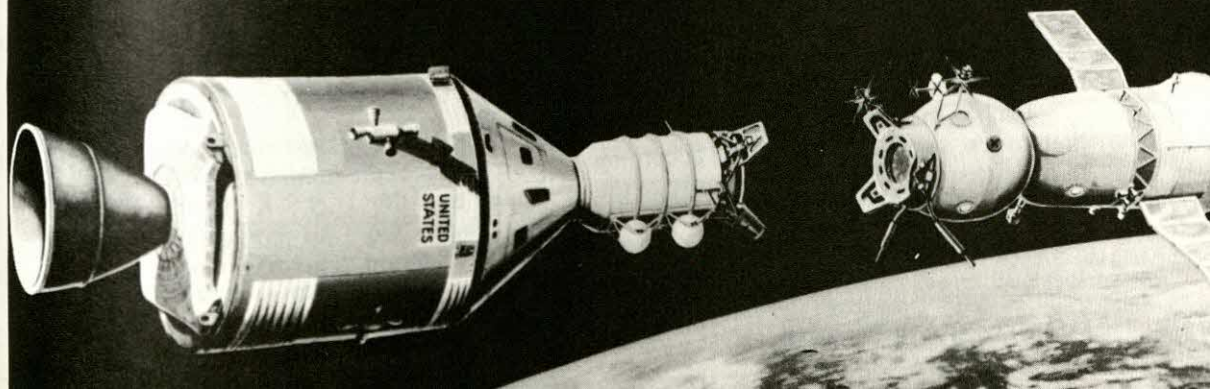
Collisions with meteorites. The meteorites seen in museums, ranging in size from fair-sized rocks to large boulders, represent exceptionally rare specimens. Scientists have known for many years that the great majority of the meteorites that strike our atmosphere are exceedingly small. Tests carried out with unmanned satellites showed that the danger of damaging collisions with meteorites was slight. The ordinary meteorites encountered by craft in space would be so tiny that they could not possibly damage the walls of the craft. It was calculated that a spacecraft might remain in space for years without encountering a meteorite larger than a few centimeters in diameter.

Reentry into the earth's atmosphere. It was necessary to protect manned spacecraft from the tremendous heat that would be produced as the craft would plunge through the atmosphere on its return from outer space. The technique finally devised consisted of slowing down the craft by braking rockets, called retro-rockets, and by providing the craft with a shield to absorb the heat produced during the period of reentry.

Other problems of manned space flight. Investigators realized that it would be

NASA

An artist's conception of a crucial part of the Apollo-Soyuz mission. The U.S. Apollo craft has already linked with the docking module, and is approaching the U.S.S.R.'s Soyuz craft for a docking maneuver.



necessary to provide adequate amounts of oxygen for breathing purposes and to dispose of the carbon dioxide and water vapor produced in exhalation. Effective methods of taking food and drink would have to be developed, since in a state of weightlessness, ordinary techniques could not be employed. It would also be necessary to maintain an endurable temperature within the spacecraft.

MANNED CAPSULE IN FLIGHT

All spacecraft, manned or unmanned, must be launched into space by multistage rockets. A single-stage rocket cannot achieve the necessary velocity. As a precautionary measure, each manned capsule has, attached to its nose, an escape tower provided with its own rocket engines. If there appears to be danger of fire or explosion at the end of a launch, the rocket engines of the tower can be actuated, either by the range safety officer or by an astronaut in the capsule. This causes the capsule to pull away from the booster rockets. After it reaches the proper distance from the landing site, a parachute landing can be made. If the launching goes well, the tower is later discarded.

The capsule is in the shape of a cone with an apex lopped off. In U.S. craft, the atmosphere within the capsule is made up of oxygen, pumped from tanks. For extended flights, a two-gas system—oxygen plus nitrogen or helium—is used. Such a system has been used by the Soviet Union in their Soyuz spacecraft.

The capsule has an effective ventilation system. Air is continuously drawn from the interior by a fan. It passes through a filter, which removes the carbon dioxide exhaled by the astronauts; through an evaporator, which absorbs the water vapor resulting from exhalation; and through a heat exchanger, where it is water-cooled. The purified and cooled air is then directed back to the interior of the capsule.

Each astronaut is provided with a contour couch, specially constructed according to his measurements. Since the nose of the capsule points straight upward at the moment of launch, the astronaut is in the su-

pine position. His space suit is connected by means of piping to a different system of tanks and pumps from that servicing the capsule. Once in orbit, the astronaut can remove his helmet and breathe the oxygen in the capsule. If anything were to go wrong with the capsule's ventilating and filtering system, the astronaut would replace his helmet, and the space-suit system would take over.

Each capsule has an elaborate control system, based on small rocket engines set around the craft. Controls are automatic during the launch. Once in orbit, the pilot takes over, though the controls can be set at automatic if so desired. Obviously the pilot takes over the controls in crucial maneuvers such as rendezvous in space, space docking, and taking the craft out of orbit.

None of the maneuvers are more vital than those required to bring about safe reentry through the atmosphere. When the time for reentry is at hand, the command pilot turns his craft around so that it is moving with the blunt end—the base of the cone—foremost. Retro-rockets mounted in this blunt end are then fired, decreasing the speed of the craft. At the same time, a shield in the craft's blunt end absorbs much of the heat of friction as it collides with more and more atmospheric particles.

All U.S. space-flight landings, with the exception of those of the space shuttle, have taken place at sea. For water landings, the capsule's speed is reduced to a few hundred kilometers per hour, and when the craft is reasonably near the target area, small parachutes, called *drogues*, open, stabilizing the craft and further decreasing its speed. Finally, when the capsule is about 3,000 meters above the target area, the large main parachutes open. The velocity is reduced to about 35 kilometers per hour, and the craft makes a gentle landing upon the sea.

THE MOON-EXPLORATION PROGRAM

Since 1961, numerous manned spacecraft have been launched by the United States and the U.S.S.R. Until Skylab 1, U.S. flights were all part of a program to reach the moon. The decision to make the

MANNED SPACE FLIGHTS

| Date of flight | Name of spacecraft | Name(s) of astronaut(s) or cosmonaut(s) | No. of orbits | Duration of flight | Remarks |
|-------------------|---------------------------------|---|---------------|--------------------|--|
| Apr. 12, 1961 | <i>Vostok 1</i> (U.S.S.R.) | Yuri A. Gagarin | 1 | 1.8 hrs. | First manned space flight |
| May 5, 1961 | <i>Freedom 7</i> (U.S.) | Alan B. Shepard, Jr. | — | 15 min. | First American in space |
| July 21, 1961 | <i>Liberty Bell 7</i> (U.S.) | Virgil I. Grissom | — | 16 min. | |
| Aug. 6-7, 1961 | <i>Vostok 2</i> (U.S.S.R.) | Gherman S. Titov | 17 | 25.3 hrs. | |
| Feb. 20, 1962 | <i>Friendship 7</i> (U.S.) | John H. Glenn, Jr. | 3 | 4.9 hrs. | First U.S. manned orbital mission |
| May 24, 1962 | <i>Aurora 7</i> (U.S.) | M. Scott Carpenter | 3 | 4.9 hrs. | |
| Aug. 11-15, 1962 | <i>Vostok 3</i> (U.S.S.R.) | Andrian G. Nikolayev | 64 | 94.3 hrs. | Part of first Soviet "group flight" |
| Aug. 12-15, 1962 | <i>Vostok 4</i> (U.S.S.R.) | Pavel R. Popovich | 48 | 71.0 hrs. | Came within 4.9 kilometers of <i>Vostok 3</i> on first orbit |
| Oct. 3, 1962 | <i>Sigma 7</i> (U.S.) | Walter M. Schirra, Jr. | 6 | 9.2 hrs. | |
| May 15-16, 1963 | <i>Faith 7</i> (U.S.) | L. Gordon Cooper, Jr. | 22 | 34.3 hrs. | Last flight of Mercury program |
| June 14-19, 1963 | <i>Vostok 5</i> (U.S.S.R.) | Valery F. Bykovsky | 81 | 119.1 hrs. | Part of second Soviet "group flight" |
| June 16-19, 1963 | <i>Vostok 6</i> (U.S.S.R.) | Valentina Tereshkova | 48 | 70.8 hrs. | Came within 4.8 kilometers of <i>Vostok 5</i> . First woman in space |
| Oct. 12-13, 1964 | <i>Voskhod 1</i> (U.S.S.R.) | Vladimir M. Komarov Konstantin P. Feoktistov Boris B. Yegorov | 16 | 24.3 hrs. | First 3-man crew in space |
| Mar. 18-19, 1965 | <i>Voskhod 2</i> (U.S.S.R.) | Pavel I. Belyayev Alekssei A. Leonov | 17 | 26.0 hrs. | Leonov spent 10 min. outside spacecraft; first "walk in space" |
| Mar. 23, 1965 | <i>Gemini 3</i> (U.S.) | Virgil I. Grissom John W. Young | 3 | 4.9 hrs. | First U.S. 2-man crew in space |
| June 3-7, 1965 | <i>Gemini 4</i> (U.S.) | James A. McDivitt Edward H. White | 62 | 97.9 hrs. | White spent 21 min. outside spacecraft |
| Aug. 21-29, 1965 | <i>Gemini 5</i> (U.S.) | L. Gordon Cooper, Jr. Charles Conrad, Jr. | 120 | 190.9 hrs. | First extended U.S. manned flight |
| Dec. 4-18, 1965 | <i>Gemini 7</i> (U.S.) | Frank Borman James A. Lovell | 206 | 330.6 hrs. | Served as rendezvous target for <i>Gemini 6</i> |
| Dec. 15-16, 1965 | <i>Gemini 6</i> (U.S.) | Walter M. Schirra, Jr. Thomas P. Stafford | 16 | 25.9 hrs. | Rendezvoused within 0.3 meters of <i>Gemini 7</i> |
| Mar. 16, 1966 | <i>Gemini 8</i> (U.S.) | David R. Scott Neil A. Armstrong | 7 | 10.7 hrs. | First "docking" in space. Docked with unmanned <i>Agena</i> target vehicle. <i>Gemini 8</i> was forced down because of failure of maneuvering rocket |
| June 3-6, 1966 | <i>Gemini 9</i> (U.S.) | Thomas P. Stafford Eugene A. Cernan | 44 | 72.3 hrs. | Made triple rendezvous with orbiting target vehicle; Cernan spent 2 hrs. 5 min. outside spacecraft |
| July 18-21, 1966 | <i>Gemini 10</i> (U.S.) | Michael Collins John W. Young | 43 | 70.8 hrs. | Linked in space with orbiting <i>Agena</i> target; when latter's rockets fired, both vehicles reached record height of 764 kilometers. <i>Gemini</i> also rendezvoused with other <i>Agena</i> craft |
| Sept. 12-15, 1966 | <i>Gemini 11</i> (U.S.) | Charles Conrad, Jr. Richard F. Gordon, Jr. | 44 | 71.3 hrs. | Docked with <i>Agena</i> target; reached record height of 1,370 kilometers; Gordon made brief space walk |

MANNED SPACE FLIGHTS

| Date of flight | Name of spacecraft | Name(s) of astronaut(s) or cosmonaut(s)* | No. of orbits | Duration of flight | Remarks |
|----------------------|----------------------------|--|--------------------------------|--------------------|--|
| Nov. 11-15, 1966 | <i>Gemini 12</i> (U.S.) | James A. Lovell, Jr. Edwin E. Aldrin, Jr. | 59 | 94.6 hrs. | Docked with <i>Agena</i> target vehicle; Aldrin engaged in extra-vehicular activity for more than 5½ hours |
| Apr. 23-24, 1967 | <i>Soyuz 1</i> (U.S.S.R.) | Vladimir M. Komarov | 17 | 26.7 hrs. | Komarov killed when vehicle crashed; first actual space fatality |
| Oct. 11-22, 1968 | <i>Apollo 7</i> (U.S.) | Walter M. Schirra, Jr. Donn F. Eisele R. Walter Cunningham | 163 | 260.1 hrs. | Successful flight test of three-man <i>Apollo</i> command module |
| Oct. 26-30, 1968 | <i>Soyuz 3</i> (U.S.S.R.) | Georgi T. Beragovoi | 60 | 94.8 hrs. | Rendezvous with unmanned <i>Soyuz 2</i> |
| Dec. 21-27, 1968 | <i>Apollo 8</i> (U.S.) | Frank Borman James A. Lovell, Jr. William A. Anders | 10 (around moon) | 147 hrs. | First manned flight around the moon |
| Jan. 14-17, 1969 | <i>Soyuz 4</i> (U.S.S.R.) | Vladimir A. Shatalov | 45 | 71.2 hrs. | Rendezvous with <i>Soyuz 5</i> |
| Jan. 15-18, 1969 | <i>Soyuz 5</i> (U.S.S.R.) | Boris V. Volynov Aleksei S. Yeliseyev Yevgeni V. Khrunov | 46 | 72.7 hrs. | Two men transferred to <i>Soyuz 4</i> and landed with it |
| Mar. 3-13, 1969 | <i>Apollo 9</i> (U.S.) | James A. McDivitt David R. Scott Russell L. Schweikart | 151 | 241 hrs. | First docking with lunar module |
| May 18-26, 1969 | <i>Apollo 10</i> (U.S.) | Thomas P. Stafford Eugene A. Cernan John W. Young | 31 (around moon) | 192 hrs. | Descent in lunar module to within 14.5 kilometers of moon |
| July 16-24, 1969 | <i>Apollo 11</i> (U.S.) | Neil A. Armstrong Edwin E. Aldrin, Jr. Michael Collins | | 195 hrs. | First manned landing on the moon; Mare Tranquillitatis; lunar EVA time totalled 2 hrs. 13 min. |
| Oct. 11-16, 1969 | <i>Soyuz 6</i> (U.S.S.R.) | Georgi S. Shonin Valery N. Kubasov | 80 | 118.7 hrs. | Together with <i>Soyuz 7</i> and <i>Soyuz 8</i> , world's first triple launch of manned vehicles; first welding experiments in space |
| Oct. 12-17, 1969 | <i>Soyuz 7</i> (U.S.S.R.) | Anatoly V. Filipchenko Viktor V. Gorbato Vladislav N. Volkov | 80 | 118.7 hrs. | Rendezvous maneuvers with <i>Soyuz 8</i> |
| Oct. 13-18, 1969 | <i>Soyuz 8</i> (U.S.S.R.) | Vladimir A. Shatalov Aleksei S. Yeliseyev | 80 | 118.7 hrs. | |
| Nov. 14-24, 1969 | <i>Apollo 12</i> (U.S.) | Charles Conrad, Jr. Richard F. Gordon, Jr. Alan L. Bean | | 244.6 hrs. | Second manned lunar landing; Oceanus Procellarum; two lunar EVA's totalling 7 hrs. 39 min. |
| Apr. 11-17, 1970 | <i>Apollo 13</i> (U.S.) | James A. Lovell, Jr. Fred W. Haise, Jr. John L. Swigert, Jr. | — | 142.9 hrs. | Planned lunar landing aborted after rupture of oxygen tank in service module |
| June 2-19, 1970 | <i>Soyuz 9</i> (U.S.S.R.) | Andrian G. Nikolayev Vitaly I. Sevastyanov | 287 | 425 hrs. | Tested man's reactions to long periods of weightlessness |
| Jan. 31-Feb. 9, 1971 | <i>Apollo 14</i> (U.S.) | Alan B. Shepard, Jr. Stuart A. Roosa Edgar D. Mitchell | | 216.7 hrs. | Third manned lunar landing; Fra Mauro highlands; two lunar EVA's totalling 9 hrs. 19 min. |
| Apr. 23-25, 1971 | <i>Soyuz 10</i> (U.S.S.R.) | Vladimir A. Shatalov Nikolai N. Rukavishnikov Aleksei S. Yeliseyev | 32 | 59.8 hrs. | Docking with <i>Salyut</i> orbital space station |
| June 6-30, 1971 | <i>Soyuz 11</i> (U.S.S.R.) | Georgi T. Dobrovolsky Vladislav N. Volkov Viktor I. Patsayev | 359 (including <i>Salyut</i>) | 569.7 hrs. | Longest stay in space; rendezvous with <i>Salyut</i> space station, which cosmonauts occupy; all three die of accidental depressurization as <i>Soyuz</i> returns to earth |
| July 26-Aug. 7, 1971 | <i>Apollo 15</i> (U.S.) | David R. Scott James B. Irwin Alfred A. Worden | | 295.2 hrs. | Fourth manned lunar landing; Hadley Rill; three lunar EVA's for 18 hrs. 37 min.; use of lunar roving vehicle |
| Apr. 16-27, 1972 | <i>Apollo 16</i> (U.S.) | John W. Young Charles M. Duke, Jr. Thomas K. Mattingly | | 319.8 hrs. | Fifth manned lunar landing; Descartes crater; three lunar EVA's for 20 hrs. 14 min.; use of LRV; lunar orbiter launched |

MANNED SPACE FLIGHTS

| Date of flight | Name of spacecraft | Name(s) of astronaut(s) or cosmonaut(s)* | No. of orbits | Duration of flight | Remarks |
|----------------------------|----------------------------|---|---------------|-------------------------|--|
| Dec. 7–19, 1972 | <i>Apollo 17</i> (U.S.) | Eugene Cernan Harrison Schmitt Ronald E. Evans | | 301.8 hrs. | Sixth and last U.S. manned lunar landing; Taurus-Littrow Valley. |
| May 25–June 22, 1973 | <i>Skylab I</i> (U.S.) | Charles Conrad, Jr. Joseph P. Kerwin Paul J. Weitz | 395 | 672.7 hrs. | Manned earth-orbiting space station; test of human space endurance; studies of space, earth. |
| July 28–Sept. 26, 1973 | <i>Skylab II</i> (U.S.) | Alan L. Bean Jack R. Lousma Owen K. Garriott | 859 | 1,427.0 hrs. | Manned earth-orbiting space station; record test of human space endurance. |
| Nov. 16, 1973–Feb. 8, 1974 | <i>Skylab III</i> (U.S.) | Gerald P. Carr Edward G. Gibson William R. Pogue | | 84 days, 1 hr., 16 min. | Manned earth-orbiting space station; record test of human space endurance; studies of sun, space, earth resources. |
| Dec. 18–26, 1973 | <i>Soyuz 13</i> (U.S.S.R.) | Pyotr Klimuk Valentin Lebedev | | 9 days | Test of spacecraft that serves <i>Salyut</i> space station. |
| July 3–19, 1974 | <i>Soyuz 14</i> (U.S.S.R.) | Pavel Popovich Yuri Artyukhin | | 17 days | Docking with <i>Salyut 3</i> and crew transfer to it. |
| Jan. 11–Feb. 9, 1975 | <i>Soyuz 17</i> (U.S.S.R.) | Aleksei Gubarev Georgi Grechko | | 30 days | Docking with <i>Salyut 4</i> space station; crew spends 28 days in <i>Salyut</i> ; studies of sun, stars. |
| May 24–July 26, 1975 | <i>Soyuz 18</i> (U.S.S.R.) | Pyotr Klimuk Vitaly Sevastyanov | | 63 days | Docks with <i>Salyut 4</i> space station; experiments involving sun, outer space. |
| July 15–24, 1975 | <i>Apollo</i> (U.S.) | Thomas P. Stafford Donald K. Slayton | 136 | 9 days 6 days | <i>Apollo-Soyuz Test Project</i> , the first cooperative international space flight. <i>Apollo</i> successfully docks with <i>Soyuz</i> . |
| July 15–21, 1975 | <i>Soyuz 19</i> (U.S.S.R.) | Vance D. Brand Aleksei A. Leonov Valery N. Kubasov | | | |
| Feb. 7–25, 1977 | <i>Soyuz 24</i> (U.S.S.R.) | Viktor V. Gorbatko Yuri Glazkov | | 18 days | Research on the effects of weightlessness. |
| April 9–Oct. 11, 1980 | <i>Soyuz 35</i> (U.S.S.R.) | Leonid Popov Valery Ryumin | | 185 days | Docked with <i>Salyut 6</i> for scientific research and repairs to space station. |
| Mar. 13–May 26, 1981 | <i>Soyuz 38</i> (U.S.S.R.) | Vladimir Kovalyonok Viktor Savinykh | | 75 days | Docked with <i>Salyut 6</i> ; made repairs; had visitors via <i>Soyuz 39</i> & <i>40</i> before returning to earth. <i>Salyut</i> station then left empty. |
| April 12–14, 1981 | <i>Columbia 1</i> (U.S.) | John W. Young Robert J. Crippen | | 54.5 hrs. | First flight of a reusable space vehicle (the shuttle). |
| Nov. 12–14, 1981 | <i>Columbia 2</i> (U.S.) | Joe H. Engle Richard H. Truly | | 54 hrs. | Tested robot arm; used special radar to study earth. |
| Mar. 22–30, 1982 | <i>Columbia 3</i> (U.S.) | Jack R. Lousma C. Gordon Fullerton | | 8 days | Manipulated robot arm; research in physics, biology. |
| June 27–July 4, 1982 | <i>Columbia 4</i> (U.S.) | Thomas K. Mattingly, 2d Henry W. Hartsfield, Jr. | | 8 days | Final test of shuttle's operational readiness; first landing on a hard-surfaced runway. |
| Nov. 11–16, 1982 | <i>Columbia 5</i> (U.S.) | Vance D. Brand Robert F. Overmyer William B. Lenoir Joseph P. Allen | | 6 days | First operational shuttle mission; launched two communications satellites; planned EVA canceled because of space-suit malfunction. |
| May 13–Dec. 10, 1982 | <i>Soyuz 41</i> (U.S.S.R.) | Anatoly Berezhovoy Valentin Lebedev | | 211 days | First to dock with new <i>Salyut 7</i> space station; visited by two space teams before return to earth. |
| Apr. 4–9, 1983 | <i>Challenger 1</i> (U.S.) | Paul J. Weitz Karol J. Bobko Story Musgrave Donald H. Peterson | | 5 days | Released tracking and data relay satellite; practiced two-man EVA's for 3 hrs., 52 min.; tested minimum landing distance. |
| June 18–24, 1983 | <i>Challenger 2</i> (U.S.) | Robert L. Crippen Frederick H. Hauck John M. Fabian Sally K. Ride Norman E. Thagard | | 6 days | Released Canadian and Indonesian communications satellites; tested 50-ft mechanical arm; carried the first American woman mission specialist into space. |
| Aug. 30–Sept. 5, 1983 | <i>Challenger 3</i> (U.S.) | Richard H. Truly Daniel C. Bradenstein Guion S. Bluford, Jr. Dale Gardner William E. Thornton | | 6 days | Released communications and weather satellite for India; performed first nighttime launch and nighttime landing; carried first U.S. black astronaut (Bluford). |

moon the first objective was an obvious one, since it is only 385,000 kilometers distant—much nearer to the earth than any other celestial body.

The U.S. moon-exploration program was based on three successive stages: Projects Mercury, Gemini, and Apollo. Project Mercury, which was completed in 1963, represented the preliminary phase. Spacecraft with only one astronaut were sent aloft. The program was intended to test the reactions of the pilot and his ability to perform various tasks while in a state of weightlessness. It was also designed to solve basic problems of space flight—including reentry into the earth's atmosphere.

Project Gemini was completed in late 1966. Each spacecraft contained two astronauts who carried on various tasks connected with the project of landing a man on the moon. They also executed various vital maneuvers, including taking the craft off course. They practiced rendezvous in space, with two capsules meeting at a designated point in the heavens. Likewise they carried out space docking: connecting two craft in space. Among the most spectacular feats accomplished in the course of Project Gemini were the "space walks." In a space walk, one of the astronauts left the capsule and, tethered to it by a lifeline, "walked in space." This was referred to as EVA, or extravehicular activity. The lifeline used by these astronauts was called an umbilical line. It contained piping for oxygen supplied by the capsule's pumping system and electric wiring for communication between the "spacewalker" and the astronaut in the capsule.

The aim of Project Apollo, the third stage of the moon-exploration program, was to land men on the moon. Three series of unmanned probes were a preliminary part of the Apollo program: the Ranger, Lunar Orbiter, and Surveyor series. The Ranger Project telecast to earth thousands of closeups of the lunar surface. For the first time, features as small as 25 centimeters across were visible to man. The Lunar Orbiter series involved five Orbiters, launched between August 1966 and August 1967. These spacecraft orbited the moon.

They provided clear pictures of most of the lunar surface, including the side never seen from the earth. Shots of the earth as seen from the vicinity of the moon were also taken by Orbiter 5. Between June 1966 and January 1968, the Surveyor program succeeded in soft-landing five craft on the moon. These probes took thousands of pictures of the lunar surface and sky, including an eclipse of the sun by the earth. Beginning with Surveyor 3, the crafts had mechanically-operated shovels that scooped up and analyzed the lunar soil. Data provided by the Orbiter and Surveyor series helped scientists select landing sites for Apollo missions.

The early Apollo spacecraft were launched by the Saturn IB, a two-stage rocket with an initial thrust of 750,000 kilograms. Beginning with Apollo 8, the launch vehicle has been the Saturn V rocket. This powerful three-stage vehicle is 86 meters high and has a takeoff thrust of about 3,500,000 kilograms.

The spacecraft consists of three basic parts: the command module (CM), the service module (SM), and the lunar module (LM). In addition, there is a launch escape system, which can thrust the command module containing the astronauts to safety in case of a malfunction during the initial launching stages. The Apollo is attached to the Saturn V by the spacecraft-lunar module adapter.

The command module is the control center of the spacecraft. It is the working and living area for the three astronauts during the mission (except for the period when two of the men are in the lunar module and on the moon). The service module houses electrical and propulsion systems as well as most of the spacecraft's oxygen supply. The lunar module is employed for the actual exploration of the moon's surface. It consists of two parts: the ascent stage and the descent stage.

The initial manned flights of the Apollo spacecraft were to occur in 1967. However, in January 1967, while testing their Apollo vehicle at Cape Kennedy, three astronauts died. The capsule caught fire when uncovered wiring caused a spark in the

Ground-based test crews conducting biomedical experiments in a Skylab mockup. Such tests provide information on how man reacts to the conditions of space travel.

pure-oxygen atmosphere. The ensuing investigations and redesign of the vehicle delayed the program for more than a year. In early 1968 two unmanned Apollo flights were launched. Then in October 1968 Apollo 7 orbited the earth. During the 11-day mission the astronauts flight-tested their vehicle, performed intricate rendezvous maneuvers, and relayed live-television pictures "from the lovely Apollo room, high atop everything."

Apollo 8 sent back even more exciting telecasts. Hundreds of millions of television viewers around the world saw pictures of lunar mountains and craters, the lunar terminator and "the good earth." This was man's first trip to the vicinity of the moon. As the craft orbited the moon, the astronauts photographed and studied the lunar



NASA

Astronaut Ronald E. Evans was the command module pilot on the flight of Apollo 17. During the homeward journey, Evans took a space walk to recover some photographic equipment mounted on the craft's exterior.

NASA



Moments after a spacecraft splashes down, para-rescuemen jump into the ocean from a helicopter hovering near the capsule. They carry a life raft that will be inflated alongside the capsule.



NASA

surface, paying particular attention to areas considered for future landing sites.

The lunar module received its first space test during the earth-orbital flight of Apollo 9. For the first time, men orbited the earth in a spaceship incapable of reentering the earth's atmosphere. Two of the astronauts put the LM through a series of important maneuvers simulating those that would be made in future lunar missions.

Apollo 10 was the final dress rehearsal before landing men on the moon. As the command and service modules orbited the moon, the lunar module descended to within 15 kilometers of the lunar surface. The astronauts practiced a complex series of landing maneuvers, tested the LM's landing radar, and photographed the landing site.

The near-perfect success of these missions made almost inevitable NASA's decision to proceed with the launching of Apollo 11 on July 16, 1969.

The epic voyage of Apollo 11 landed Neil Armstrong and Edwin Aldrin on the surface of the moon. A third astronaut, Michael Collins, remained with the orbiting command ship during the lunar exploration. For more than two hours they walked on the lunar surface, collecting rocks, photographing the area, and setting up scientific experiments. The men moved easily in the low gravitational field, only slightly hampered by their bulky space suits. Their oxygen supplies, communications gear, and other equipment were contained in Portable Life Support Systems (PLSS) strapped to their backs.

The astronauts placed several seismometers on the lunar surface. These have enabled scientists to monitor seismic events—either moonquakes or meteorite impacts. A laser reflector that the astronauts left on the moon enabled scientists to measure the distance between the earth and the moon to an accuracy of 15 centimeters. It also provided information on the possible weakening of gravitational forces between the two bodies.

The astronauts also gathered high-velocity particles blowing out from the sun. By exposing a sheet of aluminum foil, they collected a tiny amount of this solar wind.

Of great interest to both scientists and laymen was the precious cargo of lunar rocks brought back by Apollo 11. This material weighed about 20 kilograms and included core samples as well as surface rocks and "soil." Some of the rocks were igneous, indicating the possibility of volcanic activity on the moon. The lunar soil contained a high amount of solar-wind particles.

A second collection of lunar rocks and soil was gathered by Apollo 12 astronauts Charles Conrad, Jr. and Alan Bean. Together with Richard F. Gordon, Jr., these men had lifted off the launching pad at Cape Kennedy on November 14, 1969. This time the landing site was the Ocean of Storms, 180 kilometers west of Apollo 11's Tranquility Base.

In addition to collecting geological samples, the men walked over to the Surveyor 3 probe, which had landed in this



NASA

A flotation collar has been placed around the module to stabilize it against capsizing. Later, the module will be examined and all data will be retrieved from it. The flight's three astronauts sit in the life raft with one of the para-rescuemen. The astronauts wear masks to filter out any possible lunar germs and prevent their being carried back to earth.

area in April 1967. The astronauts removed the probe's camera, scoop, and some of its tubing to take back to earth for study.

Conrad and Bean also deployed a series of instruments on the moon. These included a seismometer, a magnetometer, a solar-wind spectrometer, and lunar ionosphere and atmosphere detectors. The mission returned to earth November 24.

On April 11, 1970, Apollo 13 was launched moonward, with James A. Lovell, Jr., Fred W. Haise, Jr., and John L. Swigert, Jr., aboard. About 320,000 kilometers from earth, an oxygen-tank explosion forced the mission's return home prematurely.

The Apollo 14 mission lifted off on January 31, 1971, with Alan B. Shepard, Jr., Stuart A. Roosa, and Edgar D. Mitchell. It landed in the Fra Mauro highlands of

the moon. Shepard and Mitchell explored the area, deploying experiments and gathering 44 kilograms of rock samples. The three astronauts returned to earth February 9.

Apollo 15 left earth with David R. Scott, James B. Irwin, and Alfred M. Worden aboard, on July 26, 1971. It landed on the moon at the foot of the Apennine Mountains, near Hadley Rill. Scott and Irwin explored the area in a moon car, the LRV. They set up experiments and gathered over 77 kilograms of rock. The mission splashed down on earth August 7.

On April 16, 1972, the Apollo 16 mission was launched, with John W. Young, Charles M. Duke, Jr., and Thomas K. Mattingly. It put down on the moon near Descartes crater, in a highland region. Young and Duke made three excursions in an



NASA

The space shuttle Columbia, on its third flight, clears the Kennedy Space Center launch tower and heads for 8 days in earth orbit.

LRV and collected 97 kilograms of rock. The mission returned to earth on April 27.

On December 7, 1972, the Apollo 17 mission blasted off for the moon, with Eugene Cernan, Harrison Schmitt, and Ronald E. Evans aboard. It landed in the Taurus-Littrow Valley. Cernan and Schmitt made three trips in an LRV and collected 113 kilograms of moon rocks and soil. The mission returned on December 19.

STATIONS IN SPACE

Manned space stations have remained for long periods in orbit about the earth. The Soviet Union led the way in the 1960's with its Soyuz program. In 1971 cosmonauts began docking with the Salyut space station. A Soviet team remained aboard the Salyut 7 for 211 days in 1983.

The United States also established a space station—the Skylab—in 1973. It was equipped to house a crew of three astronauts and carried an array of instruments for studying space, the stars, the earth, and the biological effects of space on humans and other life. Three separate crews of astronauts occupied Skylab, the last in late 1973 and early 1974. Intended to stay in orbit until 1983, Skylab reentered the atmosphere in 1979 and burned.

INTERNATIONAL COOPERATION IN SPACE

In 1972 the United States and the Soviet Union agreed to embark on plans for a joint space effort. After three years of meetings, plans, practice sessions, and equipment modifications, the first joint space effort was completed in July 1975. A U.S. Apollo spacecraft, launched from Cape Canaveral, Florida, docked with a Soviet Soyuz spacecraft launched from Kazakhstan, U.S.S.R. The crafts remained linked for two days. The crews visited each other's crafts and cooperated in experiments.

THE FUTURE IN SPACE

With the successful flight of the United States space shuttle in April 1981, the U.S. space program began a new era of space technology. Reusable spacecraft, such as the shuttle, will launch and service satellites, carry out research programs, and ferry personnel and materials to and from space stations in earth orbit for various purposes.

The moon voyages, the space shuttle, and the orbiting space stations may only be the preliminaries to more extended manned exploration of space.

SPACE STATIONS

by Wernher von Braun

and Frederick I. Ordway

Space station—the words bring an image of a huge wheel-shaped structure like that often found on the covers of science fiction magazines. Such a structure was stunningly depicted as “Space Station 5” in the film *2001: A Space Odyssey*.

But space stations are not just in the imaginations of science fiction writers. The success of the United States Skylab program and the Russian Salyut program were significant steps toward the goal of space stations. The last crew to inhabit Skylab remained in orbit 84 days, setting a record for human space endurance and proving that man can live and work in space for extended periods of time.

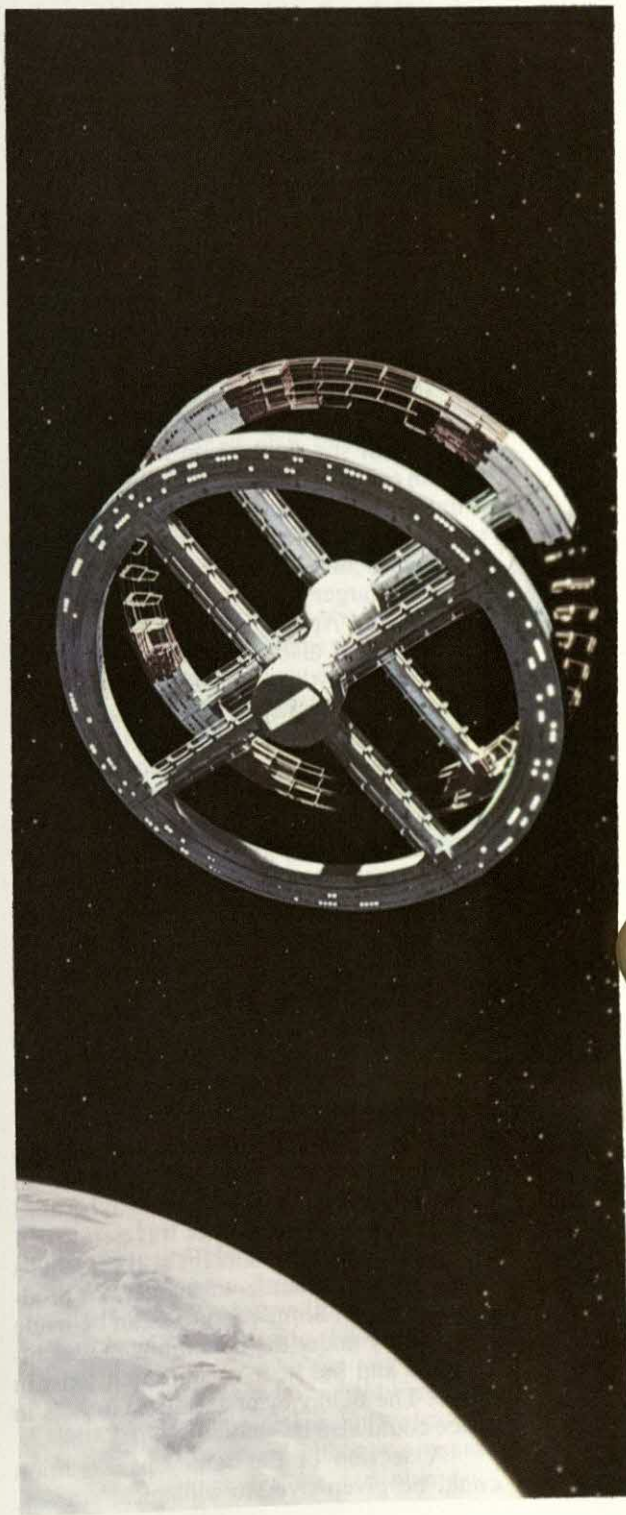
In July 1975 the successful completion of the Apollo-Soyuz Test Project was another important step. This joint Russian-American project, the first cooperative international space effort, was the result of three years of planning, meetings, practice sessions, and equipment modifications. The preliminary work—and its fruitful outcome—opened channels of communication for space scientists and engineers and allowed them to become familiar with the equipment of another country. Such steps bring the development of a practical space station and the ability to maintain it a step closer to reality.

WHAT IS A SPACE STATION?

First and foremost, a space station is an orbital spacecraft. It must be placed in orbit around the earth or some other body, such as the moon. This, of course, makes a

“Space Station 5” from the film “2001: A Space Odyssey.” The space stations of the future may well look like this.

From the MGM release, “2001: A Space Odyssey”
© 1968, Metro-Goldwyn-Mayer, Inc.



space station a kind of artificial satellite.

Second, a space station is designed, built and maintained to accommodate an astronaut crew as well as non-astronaut passengers—scientists, engineers, communications specialists and others. This makes it a large manned artificial satellite.

But a space station is still more. Space stations are larger than the manned artificial satellites we know today. They can remain in orbit for a longer time, they can carry more passengers and they are more complex.

We normally think of a space station as capable of receiving at least five or ten persons and of supplying them with what they need in order to live and work for months at a time. Ferry flights from earth would resupply the space station at intervals. These differences already indicate that a space station is larger than the familiar Mercury, Gemini, and Apollo spacecraft that orbited the earth and also larger than the Skylab "laboratory in the sky." Skylab was, incidentally, called an embryonic space station, which in many ways it was.

In discussing semipermanent stations in space, the terms space laboratory, space platform, space base and space station are all used. Generally a space laboratory is the smallest, a space station the largest.

The complexity of a space station depends on the number of persons who will live in it, on the experiments it is to perform and on the other work it is to do. A highly complex space station may be expected to perform many different kinds of astronomical and astrophysical investigations. It might also be expected to serve as a facility where medical research, particularly research on the effects of space travel on man, can be conducted. It can also be used as a space hospital for space travelers.

The space station might also serve as a space factory, manufacturing certain products on board. Some products can be made more easily under the conditions of weightlessness and the near-vacuum that exist in space. The behavior of certain materials in space could also be studied.

A section of the same space station could be given over to biological studies.

Another part of the station could house special cameras and other equipment to examine and monitor the land and oceans below, studying the earth's agricultural, mineral and other resources. Still other sensors could observe worldwide weather patterns and the atmosphere.

Someday space stations may be able to assemble, refuel and maintain spaceships traveling between planets. Truly, the potential of space stations is almost limitless.

THE BRICK MOON, AN UNUSUAL STORY

As modern and futuristic as the term space station sounds, it is an idea that has been around for a long time. More than one hundred years ago, the magazine *Atlantic Monthly* published an unusual short novel, titled *The Brick Moon*. The story appeared in four parts—in the November and December 1869 issues of the magazine and in the January and February 1870 issues. It was not until the end of 1870, however, that the author of *The Brick Moon* was revealed. It was Edward Everett Hale, the American clergyman today known to all schoolchildren as the author of the famous short story "The Man Without a Country."

The Brick Moon tells the story of a space station named the Brick Moon. The station was spherical, with a diameter of 60 meters. It was built to orbit the earth at a distance of 6,400 kilometers. It was intended to serve as a navigational or directional guidepost for oceangoing vessels.

As things turned out, the Brick Moon was launched accidentally while its construction workers and their visiting families were on board. It thus became, quite unex-

The first U.S. space station, Skylab, was successfully operated by three crews in 1973–74.

Rockwell International





Goddard Space Flight Center

A composite photo of the Middle-Atlantic section of the United States. Observation from space has provided much needed information about the earth's resources and environment.

pectedly, the world's first space station. Food and other supplies, including books, were quickly sent up to the surprised and stranded colony of 37 persons. The colonists soon became used to their new way of life. They lived in a tropical climate and were able to grow all sorts of crops.

Hale, meanwhile, said of "his little world":

"Now, let it fall. . . . Let it fall. . . .

The curve it is now on will forever clear the world . . . will forever revolve, in its obedient orbit, the Brick Moon, the blessing of all seamen—as constant in all change as its older sister [the Moon] has been fickle, and the second cynosure of all lovers upon the waves, and of all girls left behind them."

So much for the Brick Moon.

EARLY 20th CENTURY IDEAS

Many pioneering spaceflight scientists of the early 20th century showed how

space stations might be placed in orbit. They also explained the purposes that such space stations might serve. Hermann Oberth, in books published in Germany in the 1920's, suggested that space stations would be useful for communications, for refueling spaceships, and for observing the earth. If rockets were to be placed "around the Earth in a circle," he wrote, "they will behave like a small moon. Contact between them and the Earth can be maintained by means of smaller rockets."

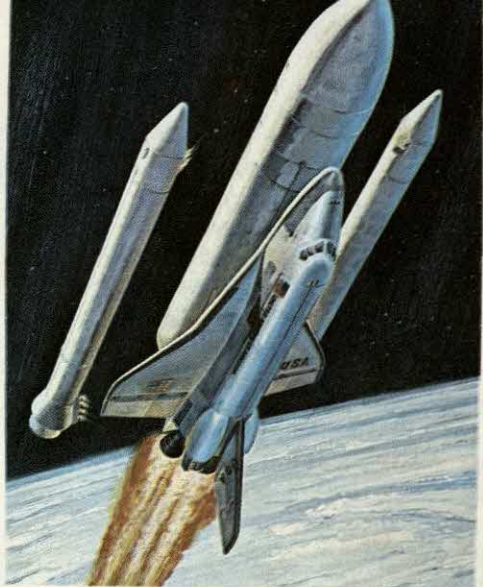
At about the same time, an Austrian Imperial Army captain, Hermann Noordung (pen name for Hermann Potochik), presented a detailed plan for a space station. He proposed a space station consisting of three main parts: a *Wohnrad*, or "living wheel," where the crew would live; a power-generating module; and an observatory.

The *Wohnrad* was shaped like a doughnut. It had a 15-meter radius and was designed to rotate, or turn, around a central hub. The rotation served to provide artificial gravity along the perimeter, or outer rim, of the doughnut.

Noordung's space station was to be 35,680 kilometers high and was to have a 24-hour orbit. Thus the space station would have the same period of rotation—24 hours—as the earth, which rotates on its axis once every 24 hours. Such an orbit is called a stationary, or geosynchronous, orbit. A space station placed in such an orbit appears stationary in the sky. Noordung thought that such an orbit would make observation of the earth below easier.

An even more ambitious idea was presented by the Englishman J. D. Bernal in his book *The World, the Flesh, and the Devil*, published in 1929. Bernal looked to the day when man would build permanent homes in space. He predicted:

"At first space navigators, and then scientists whose observations would be best conducted outside the earth, and then finally those who for any reason were dissatisfied with earthly conditions would come to inhabit [extraterrestrial] bases. Even with our present primitive knowledge we can plan out



Rockwell International

Series of illustrations showing space shuttle and some of its uses. Left: shuttle is launched by two solid propellant boosters. Right: shuttle has just jettisoned an external fuel tank.

such a celestial station in considerable detail."

Bernal's space station was spherical and extremely large—16 kilometers in diameter. He did not plan on creating artificial gravity in his station. Rather, he felt that "there is no reason to suppose that we would not ultimately adjust ourselves to weightlessness." Bernal wanted to place space stations in orbit around the sun, from where they could observe the inner solar system.

IDEAS AFTER WORLD WAR II

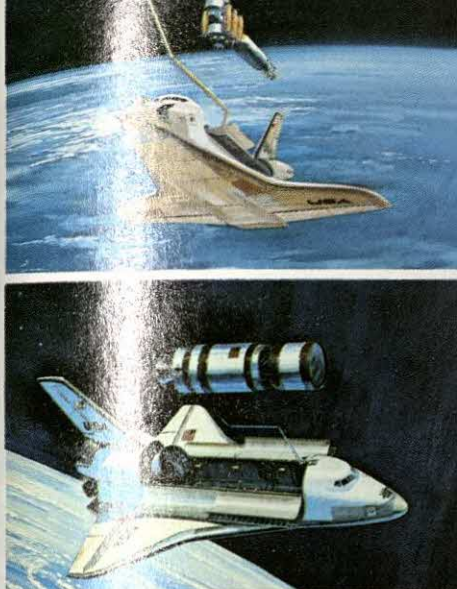
During most of the 1930's and 1940's, few original space station ideas appeared. Attention was focused on developing rocket engines and on military missiles. But, after the war, H. E. Ross of Great Britain published an article on a large rotating space station that would be used for research in meteorology, astronomy, zero-gravity conditions, high-vacuum physics, cosmic and solar radiation, and communications.

In 1951, Wernher von Braun designed a 60-meter, inflatable, wheel-shaped station to be placed in a 1,720-kilometer high orbit. He wrote that

"a person observing the earth from up there would have a unique view of cloud formation on earth, particularly above the oceans. This offers novel

possibilities for weather forecasting. By using highpowered telescopes, you may observe ships crossing the oceans and you may flash iceberg warnings to endangered ships. And, believe it or not, magnification factors could be used that would enable you to see people moving around on the earth's surface. This is because the atmospheric disturbances, when looking from outer space through the earth's atmosphere, are much less serious than those affecting astronomical observations from telescopes mounted on the bottom of the atmospheric shell. If we turn such a satellite telescope to the outer reaches of the universe, the planets and the stars, we shall find observation conditions which no terrestrial observatory could equal."

In 1953, Heinz Hermann Kölle presented his *Aussenstation* concept. It was a station that consisted of a ring made up of 36 spheres, 5 meters in diameter. The ring was connected to a central hub by 8 supporting tubes, 4 of which had elevator shafts. This station could accommodate up to 65 persons and would weigh 150 metric tons. Kölle planned his elaborate space station to serve many purposes: weather observation, forecasting and control; research on the behavior of solids, liquids, and gases in space;



Rockwell International

Left: the shuttle extends an armlike extension to retrieve a satellite and opens to deploy a satellite. Right: shuttle heads back to earth, carrying used equipment and other cargo.

biological studies of plants under zero- and low-gravity conditions; navigational aids; earth-resources observation; communications relay; and intermediate assembly, refueling and navigational assistance for spaceships.

A TIME TO RETHINK IDEAS

By the late 1940's and into the mid-1950's, engineers were thinking in terms of still larger manned space stations and, at the same time, in terms of minimal-size unmanned satellites. At this point it became clear that long before huge space stations could be built, much highly valuable research could be carried out by small unmanned satellites. The results would be radioed down to earth. Moreover, large launch vehicles needed to orbit space stations were still many years away. Therefore attention was concentrated on unmanned earth satellites. Once, however, the Sputniks, Explorers, Vanguard's, and other satellites of the late 1950's and early 1960's had been successfully orbited, interest in space stations revived.

Man would never be content to abandon the exciting field of astronautic research and exploration to remote spacecraft. Besides, there were many things man could do that unmanned craft couldn't be programmed to do.

MSOLs, MOSSs, MORLs, LORLs AND MSSs

By the mid-1960's, smaller and more realistic space-station designs began to appear. These new designs were for crews numbering no more than half a dozen to a few dozen men. Typical of the names given to them were MSOL for manned scientific orbital laboratory, MOSS for manned orbital space station, and MORL for manned orbital research laboratory.

As designed, the MORL had a useful lifetime in space of at least five years. It had two independently pressurized compartments connected by an air lock. The air lock is an airtight chamber in which air pressure can be regulated to permit transfer of people and material from an environment with one air pressure to an environment with a different pressure. The larger of the compartments was to contain a control deck from which most of the experiments would be conducted, an internal centrifuge that would create short periods of artificial gravity, and living quarters. The smaller compartment was designed as a hangar where cargo from ferry vehicles would be transferred. It was also to be used for maintenance and other support activities.

Somewhat larger space-station designs inevitably appeared—but not too much larger. One was the 24- to 36-man LORL,

or large orbital research laboratory, and another was the MSS, a 36-man multipurpose space station. MSS consisted of three experiment and crew modules, a hangar for arriving and departing ferry craft, and a zero-gravity laboratory.

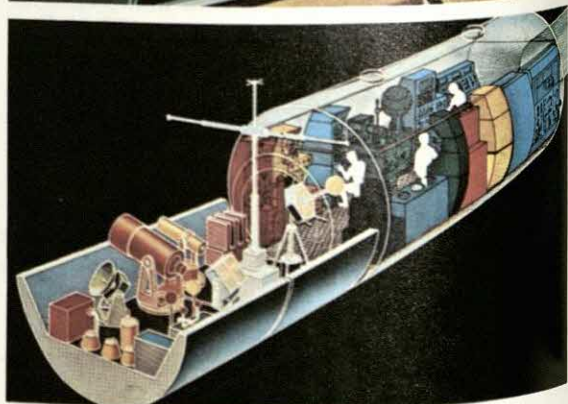
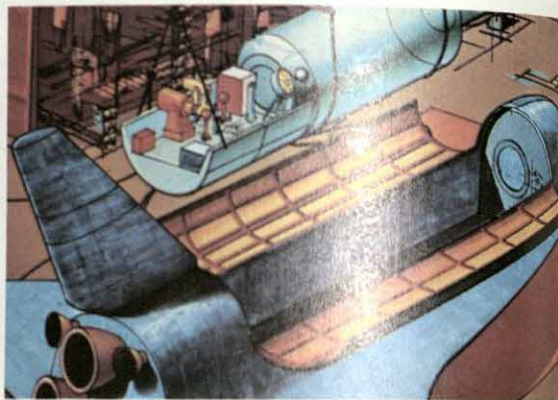
MODERN DESIGNS

Beginning in the spring of 1969, the U.S. National Aeronautics and Space Administration (NASA) and its industrial contractors concentrated their efforts on two basic types of space station. One design calls for a large, "all-in-one" *integral* station that could be placed in orbit by large launch vehicles of the Saturn 5 class. The other design is based on the *assembly-in-orbit* modular approach. Many small modules are carried aloft by space shuttles and are then assembled, or brought together, into a single unit. The core module, or main part, of an integral station, incidentally, could also receive separately orbited small modules, so, clearly, a combination of the two designs is possible.

The core module, the basic element of a 12-man integral station, would typically be a cylinder, 10 meters in diameter and 15 meters long. The cylinder would include six or seven docking ports for ferry craft, four internal decks, and a main pressurized area at each end. A central tunnel would connect the decks as well as air-lock facilities placed between the two main pressurized compartments. A tunnel would also lead to the emergency room where the astronauts would seek protection during periods of strong solar flare activity. During such periods tremendous explosions on the sun release massive electrical discharges.

The advantage of having the station divided into two pressurized compartments is that a mission could continue even if one of the compartments became damaged and had to be evacuated. Each pressure compartment would have two escape routes. All hatches would be large enough for space-suited crewmen to pass through without difficulty.

The other design—the modular, or assembly-in-orbit, design—became attractive to NASA planners once they realized



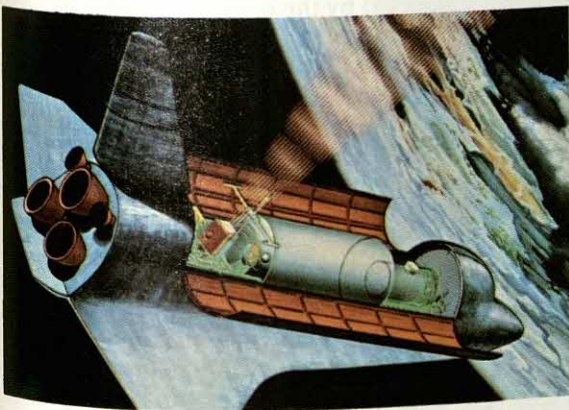
ESRO (European Space Research Organization)

that reusable space shuttles would be available by the early 1980's. The first successful shuttle of this kind, name *Columbia*, was launched and safely landed in April 1981. Shuttles will be able to carry individual passenger, cargo and experiment modules into orbit. Once in orbit these modules may be assembled into the space station proper, or be prepared for detached "free-flying" service, or simply be unloaded and made ready for return to earth. Thus, at any one time, a modular space station might consist of permanent living and experiment modules, a recently docked cargo or passenger module, and one or more temporarily docked free-flying modules.

Free-flying modules would house experiments that require extremely precise "pointing" control. The experiments must be isolated from contaminating gases and radiations produced by the station proper and by arriving and departing shuttles. All modules would be protected against micrometeoroids by "bumpers" placed over the pressurized compartments. Micrometeoroids are small solid particles traveling through space that could cause serious damage to a space station.

One advantage of a modular space station is that its individual modules can be

Series of illustrations depicting Spacelab, a project of the European Space Research Organization, (ESRO) that will be put into orbit by the Space Shuttle. Left, top to bottom: Spacelab being placed in Shuttle; member of crew moving into laboratory; crew moving into living quarters; view of external instrument platform. Below: data being transmitted to earth.



returned to earth by shuttle for repair, maintenance and modification. The shuttle can also be used to carry new supplies to the station in cargo modules. These cargo modules can return to earth with exposed film, instruments needing repairs, empty containers, worn-out equipment, trash and the like.

Rotating crews can travel in passenger modules. Should an emergency occur a shuttle can be readied, launched and be on hand at the space station in orbit within 48 hours. For even more rapid service, a special Apollo rocket could be stored in a docking port at all times, ready to carry back to earth at least part of the crew—for example, an astronaut requiring immediate medical attention on earth.

SELECTION OF ORBIT

The orbit of a space station will have to be above the earth's atmosphere, yet below the Van Allen radiation belts, a band of intense radiation surrounding the earth. Circular orbits between 465 and 510 kilometers at an angle of 50° to 55° to the equator turn out to be the best for space stations and allow the best observation of the earth.

ON BOARD THE STATION

Whether integral or modular, the space station includes a number of specialized areas in which astronauts live, work, conduct experiments, and undertake other activities. Among the areas are:

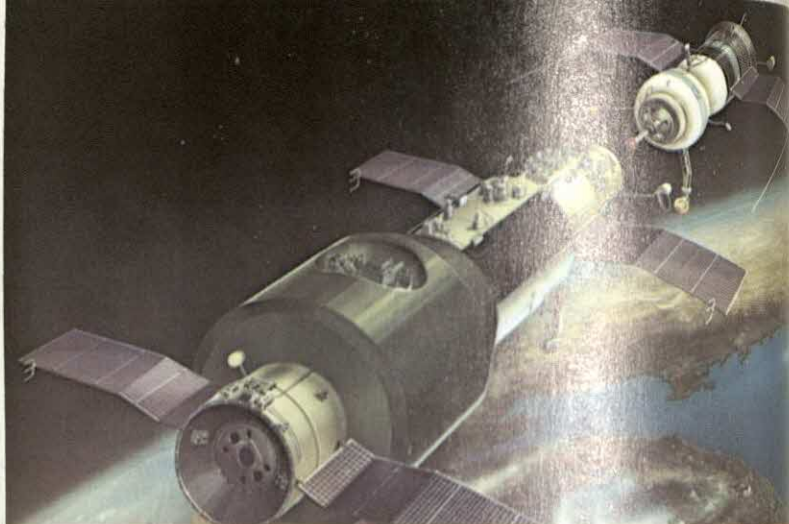
Living quarters that include sleeping rooms, toilet and washing facilities, individual study areas, clothing storage space, and medical facilities.

Wardroom and galley, where food is stored and prepared and the crew eats, relaxes, and exercises.

General-purpose area, where "everyday" types of experiments are conducted and equipment is checked over.

Command, control and data-management center to handle all communications to and from the space station. Computer systems, automatic checkout systems, and various on-board consoles and displays related to routine operations of space stations, free-flying modules and ferry craft are also used.

Model of Salyut, the first Soviet space station. Soyuz spacecraft docked with the station and crews transferred to the station.



U.S.I.S.

Storage for spare parts and miscellaneous supplies.

Provision for artificial gravity, if required.

SPACE-STATION SUBSYSTEMS

In addition to the main areas described a space station may have many subordinate systems that make it possible for the crew to carry out its many activities over long periods of time. The principal subsystems of a space station are:

Attitude, pointing and position control. These subsystems maintain the station precisely in the desired orbit and angle. They point cameras, telescopes and other sensors at a particular subject. The main elements of these subsystems are momentum wheels and an array of small rocket engines placed outside the station.

Electrical power. These subsystems provide, control and distribute electric power to the space station proper and to any free-flying modules that may be temporarily docked to the space station. The source of electric power can be either a solar-cell array or an internal nuclear reactor. Solar-cell arrays convert the rays of the sun into electric power. This system must, however, be complemented with electrochemical batteries. These batteries will provide electric power when the orbiting space station is passing through the shadow of the earth and its solar cells are not receiving any direct sunlight.

Environmental control and life support. These subsystems furnish the astronauts with an appropriate environment.

They will provide for oxygen and nitrogen gas storage and supply, for temperature and humidity control, for carbon-dioxide control, for water and waste management, for food management, and for personal hygiene.

Information handling. These subsystems display and communicate information on overall space-station operations, flight control, experiments, scheduling and many other activities. To ensure full-time contact with earth-based mission-control centers, communications from an orbiting space station will be routed through high-altitude communication satellites.

CREW OPERATIONS

In a 12-man space-station crew, three or four members will be needed to operate a typical integral or modular space station. The other eight or nine persons will thus be able to perform research and to conduct studies. Non-astronaut scientists and engineers are the "users" of the space station, and are supported by the operating crew.

USES OF SPACE STATIONS

Work in future space stations will build on research completed in February 1974 aboard the three-man Skylab orbital laboratory. Astronomers, for example, plan to conduct experiments to study X rays and other energy sources, as well as the sun and other stars. Earth-observation work will include a careful study of many of the earth's surface features; special mapping assignments; inventory of the earth's nonrenewable and renewable resources; and global

environment studies. The goals of the communications and navigation research facility are to serve international requirements for worldwide communications between ground, ocean, airborne, and spaceborne terminals. Such research should also improve ground, ocean, air, and space navigation.

Space stations will be ideal for making advanced studies in materials science and for making studies of manufacturing in the zero-gravity space environment. Scientists will try to work out on-board processing methods for new materials and products. This, in turn, may someday lead to commercial manufacturing operations in orbit.

The life sciences will be no means be neglected. A large number of space facilities and activities have been proposed to permit basic biological research to be carried out. For example, biologists are eager to study how changes in gravity affect living organisms, how differences in day-night cycles can change the functioning of an organism, how organisms age in space, and many other questions. Space biologists will also try to determine how to use the space environment to help advance medicine, agriculture, and public health.

Education will also greatly benefit from space stations. In the future, scientists and other specialists will be able to broadcast

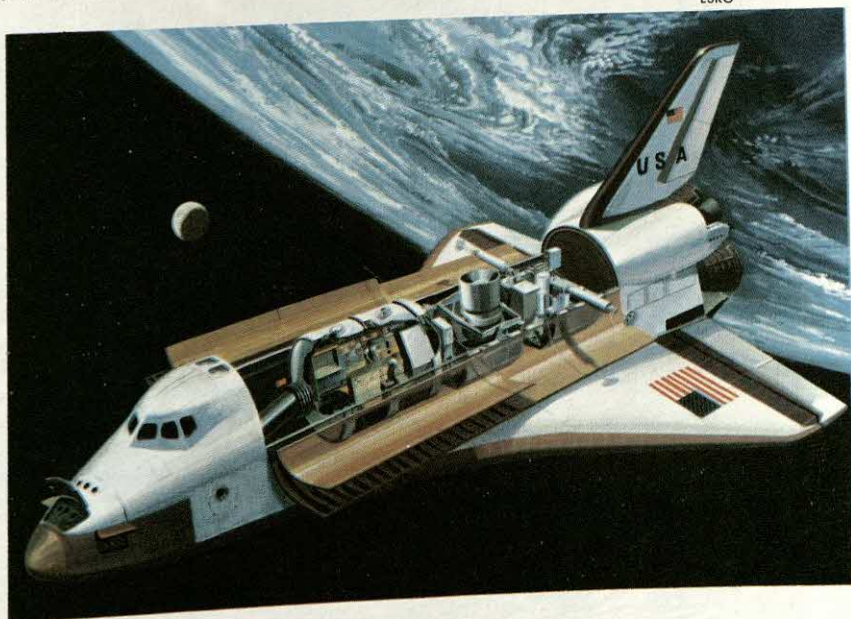
and televise summaries of their findings directly into schoolrooms all over the world.

After the establishment of a modular space station, it will be possible to move individual modules into higher orbits—including the 35,680-kilometer geosynchronous orbit—by auxiliary spacecraft called *space tugs*. Tugs contain a special propulsion system that allows them to move from one orbit to another. They can carry space-station modules all the way to orbits around the moon. Such modules, stacked up to form a two- or three-stage rocket, may also propel unmanned spacecraft from earth orbit to a Mars orbit.

In due time, 12-man stations will almost certainly be expanded, and eventually 50- to 100-man stations may be established in orbit around the earth. Since crews and researchers will most likely spend many months at a time in the space stations, the living and dining areas of the station may be provided with artificial gravity for convenience. Still further in the future may come orbital hospitals, hotels and recreational facilities. Who knows—perhaps zero-gravity gymnasts will follow up where the Skylab astronauts left off and someday perform acrobatic stunts in orbit for television audiences. Whatever can be done in space stations probably will be done.

ESRO

NASA is developing a space shuttle to ferry people and material to and from an orbiting space station. Here a model of such a shuttle with cargo doors open to reveal a model of ESRO's Spacelab.





THE EXPLORATION OF THE MOON

By Hans J. Behm

People have always been curious about the moon. Since the earliest times they have speculated about its nature and origin. For centuries the moon has also been the subject of myth and superstition. No serious attempt to understand it and to reach it could be made so long as this was the case. First, the moon had to be recognized as a distinct world in space, not unlike the earth in some respects. Second, people had to grasp the difficulties that would be involved in a possible journey to the earth's nearest neighbor.

It was the ancient Greeks who first speculated freely about the universe and who realized that the earth was a globe and that the sun, moon, and planets were perhaps worlds like the earth. The peoples of the ancient Orient had even earlier built up a tremendous store of information on the apparent motions of these heavenly lights, but they did not recognize them as distinct celestial objects. It took a long period of speculation, measurement, and study before the present state of lunar knowledge and exploration was reached.

The history of lunar study and exploration falls into three major periods: (1) the pre-telescopic era, to A.D. 1609; (2) the telescopic era, from 1609 to 1959; and (3) the space-exploration era, since 1959, the year that the first lunar probe reached the moon. In 1969 the first men set foot on the moon, an event that marked another milestone in lunar exploration.

PRE-TELESCOPIC EXPLORATION

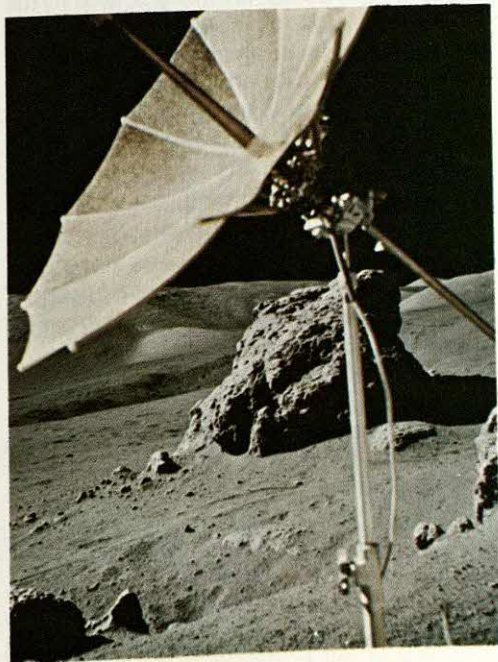
As we have stated, the ancient Greeks opened people's eyes to the possibility of worlds in space. They measured the circumference of the earth with astonishing accuracy. Their estimates of the moon's size and distance from the earth were, however, not very accurate. Lacking any

telescopes and similar instruments, they could only guess about the moon's nature. The Greeks generally believed that the earth was the center of the universe, and that the sun, moon, planets, and stars revolved around us. But the Greek astronomer, Aristarchus of Samos, in the third century B.C. proposed that the sun was the center of planetary revolutions, not the earth. This idea was generally disregarded until the sixteenth century, when it was successfully revived by the Polish astronomer, Nicolaus Copernicus.

In the absence of any exact knowledge about the atmosphere and outer space, some Greeks considered it possible to reach the moon by sailing or flight. The first

This television antenna, attached to the astronaut's Lunar Rover vehicle, enabled earth-bound scientists to explore the moon along with the astronauts.

NASA





NASA

Panoramic view of the lunar surface. Cutting across it diagonally are tracks made in the dust by a transporter device as it moved from the module (background).

known moon-voyage story was written by a Greek called Lucian of Samosata in the second century. He tells how a ship in the Atlantic is blown to the moon by a huge storm. There the passengers encounter fantastic beings who are preparing a war against the sun. Lucian wrote another story, about a man named Icarus who flies to the moon by means of bird wings attached to his body. These basically satirical tales set the pattern for many stories in the centuries that followed. The moon and other worlds were believed to be homes of manlike beings, monsters, and other weird or unusual creatures.

In the Middle Ages of Europe (A.D. 500–1500), the Catholic Church crystallized its teachings about the universe on the Ptolemaic system, which held that the earth was the center of the cosmos. Speculation about life on other worlds was considered dangerous and was discouraged. Later, however, investigations began to disclose the errors of these ideas. In 1543 Copernicus repeated Aristarchus' view that

the earth and other planets revolved around the sun. In 1609 the Italian scientist Galileo turned the newly invented telescope to the skies. What he saw there dealt a death-blow to the Ptolemaic system, and started a revolution that laid the foundations of modern astronomy.

THE TELESCOPIC ERA

The crude little telescope used by Galileo showed the face of the moon as it had never been seen before. Even its modest magnification of some thirty times showed bright mountains, craters, and dark level areas. Galileo made a crude chart of the major moon features he saw through his little telescope. He was so struck by the apparently earthlike aspect of the moon that he erroneously called the dark regions *maria*, the Latin word for "seas."

Other astronomical discoveries followed in the 1600's. Johannes Kepler, a German astronomer, announced the laws of planetary motion. One states that the planets orbit the sun in oval paths called el-



NASA

The Lunar Rover used by the Apollo 17 astronauts is dwarfed by a gigantic boulder in the mountainous Taurus-Littrow region of the moon.

lipes, and not in circles, as Copernicus had believed. The English physicist Isaac Newton developed the laws of gravitation and motion, which helped explain planetary motion. A rash of moon-voyage stories appeared, some of which incorporated the new discoveries. Most, however, were fantasies. The means of travel to the moon were totally unrealistic except for one story in which a rocket-powered vehicle was used.

As the seventeenth century progressed, astronomers become more and more aware of the vast gulfs separating the earth, moon, sun, and planets. The moon itself was recognized as a small, airless globe incapable of supporting earthly life. With the eighteenth and nineteenth centuries came tremendous progress in astronomy. Telescopes increased enormously in power and clarity, and new instruments, such as the spectroscope and the camera, vastly increased the range of knowledge. The advent of photography in the nineteenth century aided and speeded the task

of mapping the moon immeasurably. The first photograph of the moon was made by an American scientist, John Draper, in 1840. In 1897 the Paris Observatory completed a photo atlas of the moon.

By the nineteenth century, astronomy and geology had sufficiently advanced so that explanations of the moon features could be seriously offered and the origins of the features discussed.

The twentieth century has witnessed even more revolutions in astronomy. People were able to view details of the lunar surface never before seen by the human eye at the telescope. These discoveries fired small groups of people to consider seriously means for traveling to the moon. The fiction writers in the meantime had not been idle. Authors such as Jules Verne and H.G. Wells kept popular interest alive from time to time with their tales of interplanetary and lunar adventures.

Although rockets had been in use for fireworks and as weapons from the thir-

The Apollo 16 lunar module was named *Orion*. As it lifts off the moon, the flame from its ascent-stage engine creates a kaleidoscopic effect.



NASA

teenth century, few individuals ever considered them as a safe or reliable means of travel. It was not until the nineteenth century that serious proposals and even experiments were made using rocket-propelled vehicles. Near the end of the 1800's, the Russian scientist Constantin Tsiolkovsky and the German scientist Hermann Ganswindt wrote and lectured on the principles of powered space flight. But at that time, few scientists and engineers took the ideas of these two men very seriously.

The twentieth century saw the fruition of the rocket-ship dream. In the 1920's and the 1930's, the groundwork for the technology that ultimately took men and their instruments to the moon and beyond was laid. After World War II—which saw the development of rocket technology—and during the later 1940's and the 1950's, the Soviet Union and the United States sent rockets aloft to explore the upper atmosphere and near space. They then decided to launch artificial satellites to circle the earth and later to orbit or land on the moon.

SPACE EXPLORATION OF THE MOON

In 1957, the Soviet Union launched Sputnik, the first artificial satellite to orbit the earth. The United States followed suit with the Explorer 1 in early 1958. The Soviet Union then sent unmanned probes to the moon, in its series of Luniks. Later in 1959, Lunik 2 crashed on the moon. Later the same year, Lunik 3 circled the moon and sent television pictures to earth of the lunar far side, which had never before been seen

by man. It was not until 1966 that Lunik 9 made a soft landing on the moon and sent pictures of the lunar surface back to earth. In the meantime the United States planned to send unmanned probes to the moon. Its Ranger series of moon craft in 1964 and 1965 was designed to crash on the moon, but before a Ranger vehicle crashed, it took thousands of photos of the approaching surface of the moon on signal from the earth base. The photographs were transmitted to earth electronically. Three Ranger craft, 7-9, showed details of the lunar surface about a thousand times greater than could be seen with the best earth telescopes. Even small craters only one or two meters across were made visible.

From 1966 to 1967, the United States launched unmanned artificial satellites to orbit the moon. Known as Lunar Orbiters, five of these vehicles swept close to the moon's surface and took thousands of photographs, which were relayed to earth. These vehicles also took photos of the earth as seen from the moon. Another series of unmanned shots was designed to soft-land on the moon. Known as Surveyors 1-7, five of them settled gently on various spots on the moon to test its surface. They determined that the moon was safe for landing. They also carried out various tests of the moon's surface, and excavated some of it for study. Surveyor 5 made a chemical test of lunar rocks by irradiating them with atomic particles and recording how the rocks reacted. In this way scientists on earth deduced something of the

chemical composition of lunar material. They concluded it was like earthly basalt, a volcanic rock. The Surveyors showed that the moon's surface was covered with pulverized rock and larger pieces of stone. The spacecraft also sent many photos of the moon to earth, and also pictures of the earth as seen from the moon.

In the 1960's and 1970's, the Soviet Union continued its program of unmanned lunar exploration. Probes circled the moon or landed on its surface. Luna probes took samples of lunar soil and sent them, by means of rockets, back to earth for study. Two wheeled vehicles—Lunokhods 1 and 2—were landed on the moon's surface. They were driven across the lunar surface by remote control from earth.

As early as 1961, the United States had decided to land men on the moon, in Project Apollo. From late in 1968, a series of Apollo missions sent U.S. astronauts into orbit around the moon and later landed them on its surface. The astronauts brought back many specimens of rock that have immensely added to our store of knowledge about the moon. They also took thousands of photographs of the moon and the earth, on the moon itself and from the Apollo spacecraft. They set up scientific experiments on the lunar surface, the results of which were telemetered to earth.

TRAVEL ON THE MOON'S SURFACE

The moon presents many obstacles to travel on or above its surface. Conventional modes of earth transportation—airplane, automobile, train, and ship—would be very difficult or impossible there. The moon has no air. Combustion engines, such as gasoline, diesel, steam, and jet engines, cannot operate, unless a supply of air or oxygen as well as fuel is provided. The absence of an atmosphere on the moon also rules out aircraft. Only rocket or reaction power plants would make flight through the lunar skies possible. There are no water bodies for boats.

Locomotion on the lunar surface in a ground vehicle involves many problems. Motors not requiring air would be powered by electricity. The low lunar gravity—one-

sixth that at the earth's surface—would make control of the vehicle difficult. Special suspension, steering and axle systems would have to be designed.

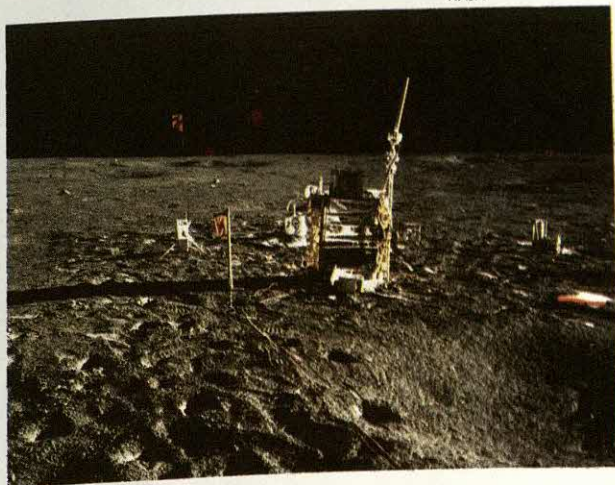
What kind of undercarriage would be suitable? Traction, or grip, on the soft and yielding lunar dust would be hard to achieve. Air-filled tires would not work. They would simply inflate and explode in the moon's vacuum. Treads or tracks, as on a tank, would be one solution. But drivers and engineers have learned that wheels with thick, flexible tires offer the best means of locomotion through soft dirt and sand. Airless tires of flexible, springlike metal or other material would be the best answer here.

Other requirements for a lunar vehicle would be lightness and compactness, for easy transport from the earth to the moon; a tough, flexible chassis and axles, for movement over rough lunar ground; an ability to withstand the vacuum conditions and great temperature extremes of the moon; and enough speed for traveling significant distances in a reasonable time. The vehicle must have room for at least two astronauts with their equipment, experimental packages, and specimens of lunar rock. Navigational and communication systems would also be vital.

The United States incorporated these features into its first lunar roving vehicle (LRV). On its first mission, it carried two U.S. astronauts on several short journeys of exploration across the moon's surface.

Central station of an Apollo Lunar Surface Experiments Package. This package contained the many devices used to study the moon.

NASA



When the aluminum LRV was unfolded from the lunar module on the moon, it was three meters in length. Each of its four wheels was powered by a one-quarter horsepower electric motor. Flexible wire-mesh tires provided excellent traction. Power was supplied by electric batteries. Top speed was about 16 kilometers per hour, maximum range, around 65 kilometers. The LRV could cross cracks in the ground 60 centimeters wide and move up slopes of 25°. Steering was accomplished by a control stick that was moved much as the one in an airplane. All four wheels could be turned, for sharp maneuvers. The car carried two astronauts and their equipment. Computerized navigation instruments told the astronauts where they were. A color television camera on the LRV relayed to earth, via a high-gain antenna, views of the moon.

The LRV's earth weight was 200 kilograms, on the moon somewhat over 34 kilograms. The vehicle could operate even if some of its vital parts broke down, which in fact did happen at first.

The mission landed in the north central region of the moon's near side, in the Palus

Putredinis ("Swamp of Decay"), between the Apennine Mountains and Mare Imbrium. In this area is a great canyonlike gash in the moon's surface—Hadley Rill, which has puzzled astronomers for generations. The astronauts made three journeys in the LRV along the Apennine mountain front and Hadley Rill, gathering many specimens of moon rock. Other LRV's were also used in the Apollo 16 and 17 missions, in 1972.

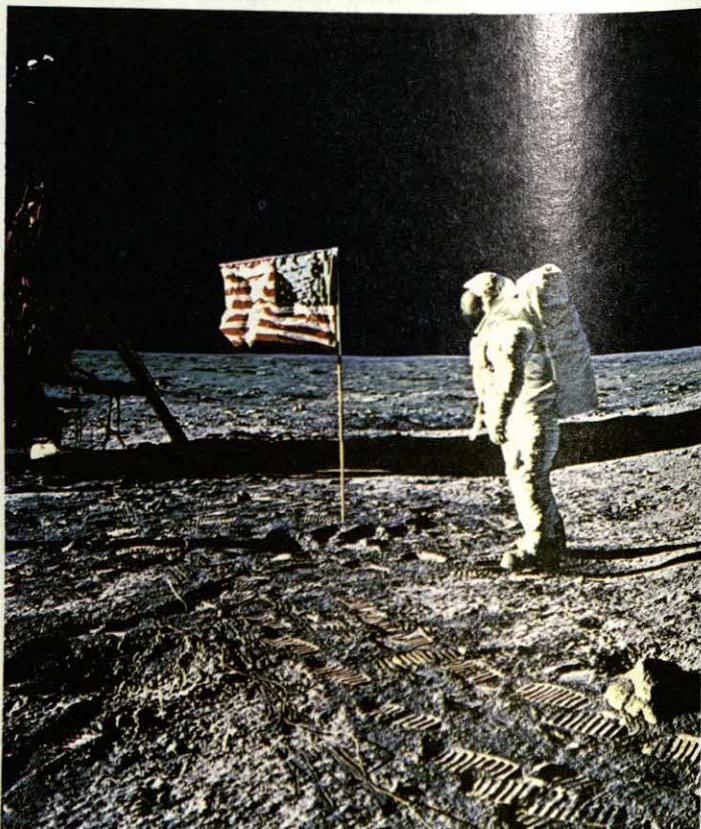
The Soviet Union has also developed lunar vehicles—Lunokhods ("Moon Rovers") 1 and 2—with several additional complications. A Lunokhod is essentially an unmanned, automated machine. This lunar vehicle is "driven" by remote control from earth. The "driver" on earth sees a television image of the lunar surface, sent by cameras on the Lunokhod as it moves.

Lunokhod itself is equipped with automated devices to control and monitor its own movements. It also carries other scientific experiments and instruments for studying the moon, inside a large container with a lid.

Lunokhod 1 was landed late in 1970 in

NASA

On July 20, 1969, Edwin E. Aldrin, Jr., the second man on the moon, poses beside the United States' flag which they placed on the moon's surface.



the Mare Imbrium, in the moon's northern hemisphere. Lunokhod 2 was set down early in 1973 in the crater Le Monnier, also in the northern hemisphere.

STUDY OF THE MOON

The field of astronomy that deals with the moon in general is called *selenology*. The branch of selenology that deals with the surface features of the moon is *selenography*. Selenology goes back to the time before telescopes. Selenography, however, begins with the telescope. In its earlier period, from 1609 into the nineteenth century, selenographers were primarily concerned with mapping the moon's surface. It was slow, painstaking work, requiring years of patient, eye-to-the-telescope observation and careful drawing.

The modern science of *astrogeology*, which deals with the structure and features of solid worlds (planetary bodies) in space outside the earth, had its origins in selenography. Other instruments besides telescopes and cameras have come to the selenographer's aid—spectroscopes, radar, and moon probes.

SURFACE FEATURES OF THE MOON

Perhaps the most controversial features of the moon are craters. Debate about

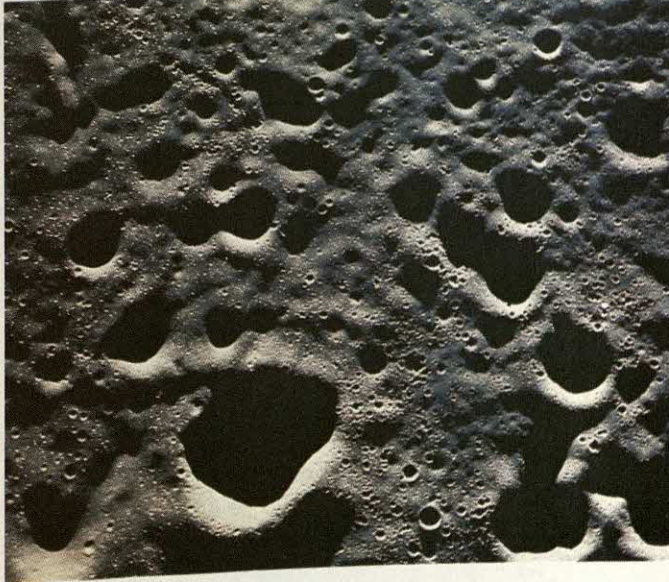
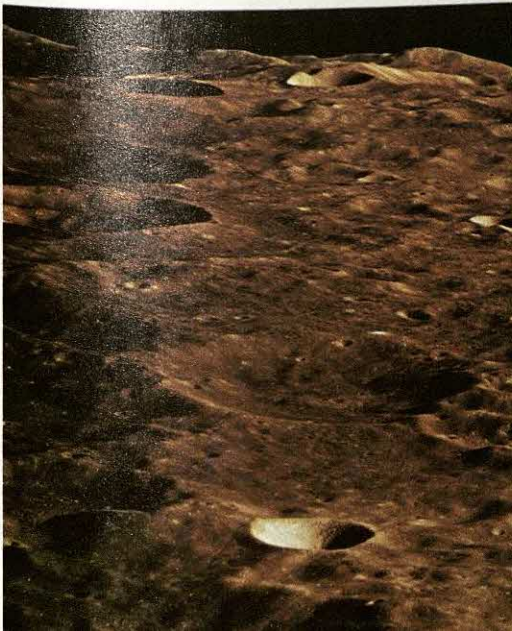
their origin still rages. Two schools of thought predominate on this point. The volcanic school holds that lunar craters were caused by eruptions of molten rock from the moon's interior. The meteoritic school contends that craters were formed by the explosive impact of meteorites from space hitting the lunar surface. Some authorities favor neither view exclusively, but feel that the craters are due to an interaction of both processes. Still others think that at least some lunar craters were created by other geologic processes, such as sinking or slumping of moon rocks, thus giving rise to circular features that superficially resemble volcanic or meteoritic craters.

Some scientists believe that meteors and even small asteroids plunging to the moon have been responsible for most of the lunar features. The impacts of huge chunks of matter from space have even been stated to originate the formation of the lunar *maria*. The *maria* tend to have a circular shape and are surrounded by mountains, much as the large craters are. The meteorists believe that the explosive impact of a meteorite affected lunar rock on a vast scale. Later, lava from the moon's interior flooded many square kilometers of the area, forming the mare.

Two of the many thousands of photos of the surface of the moon taken by the Apollo crews. Left: view from Apollo 11 shows craters which vary in width from 1 or 2 meters to 50 meters. Right: the crater-packed far side of the moon.

NASA

NASA



The volcanists argue, on the other hand, that lunar features arose by processes native to the moon. These processes are usually, but not always, much like those that shaped the features of the earth. Many scientists feel that certain surface features could not have been caused by random hits from meteorites. Volcanologists claim that the peculiar glows and hazes seen from time to time on the moon are signs of volcanic activity. Their opponents say these are luminescence in the lunar rock caused by radiation from space.

The moon probes and U.S. Apollo missions have confirmed the existence of at least some lunar features that strongly resemble volcanic features on earth—domes and lava flows. The question remains whether volcanic activity was ever widespread and important on the moon. Extensive rock melting was undoubtedly caused by impacts of large meteorites on the lunar surface. They probably also produced the large craters and the *maria*, or plains.

COMPOSITION AND HISTORY

Lunar rocks generally resemble the earth rocks known as *basalts*, which are finely grained, and *gabbros*, which are coarser-grained. These are dark, dense, chemically related rocks of igneous origin.

Among the youngest of moon rocks are the mare basalts. These rocks formed from dark lavas that poured out 3,000,000,000 to 3,800,000,000 years ago, to form the floors of many large craters and maria.

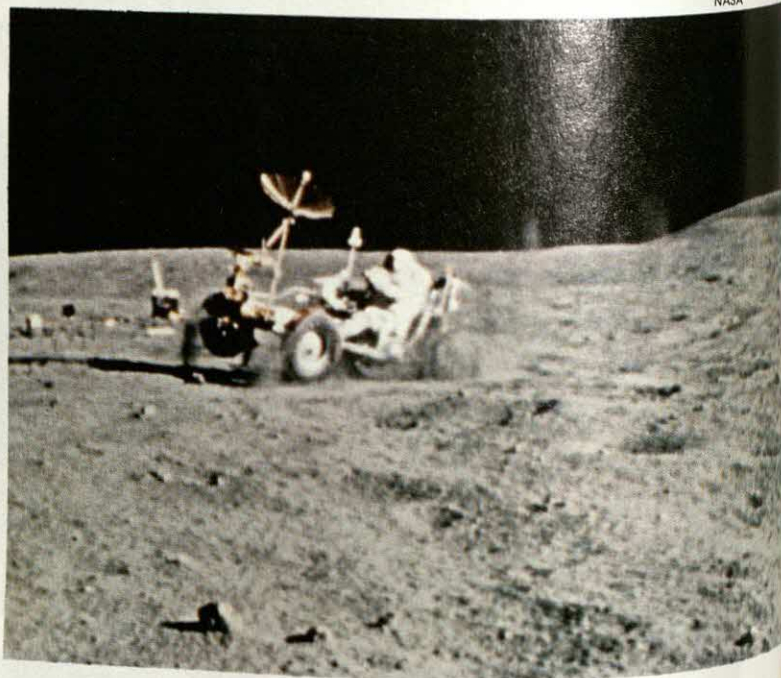
As its rocks attest, the moon has had a long, complex history, beginning about 4,600,000,000 years ago or earlier, when the original crust is believed to have solidified.

The moon has a complicated structure. There is a dense, rigid *crust*, about 24 kilometers on the earthward side and up to 64 kilometers thick on the far side. Beneath the crust is the *mantle* of denser rock, with a gabbrolite composition. There may be a partly molten layer about 965 kilometers down.

At the moon's center there may be a heavy, possibly liquid or semiliquid core about 640 kilometers in diameter. Its position is lopsided, its center closer to the earth-facing side of the moon.

The moon has a weak magnetic field. The origin of the field is still uncertain. For a more detailed discussion of the moon and of moon rocks, see the articles "The Moon" and "The Geology of the Moon" in *The New Book of Popular Science*.

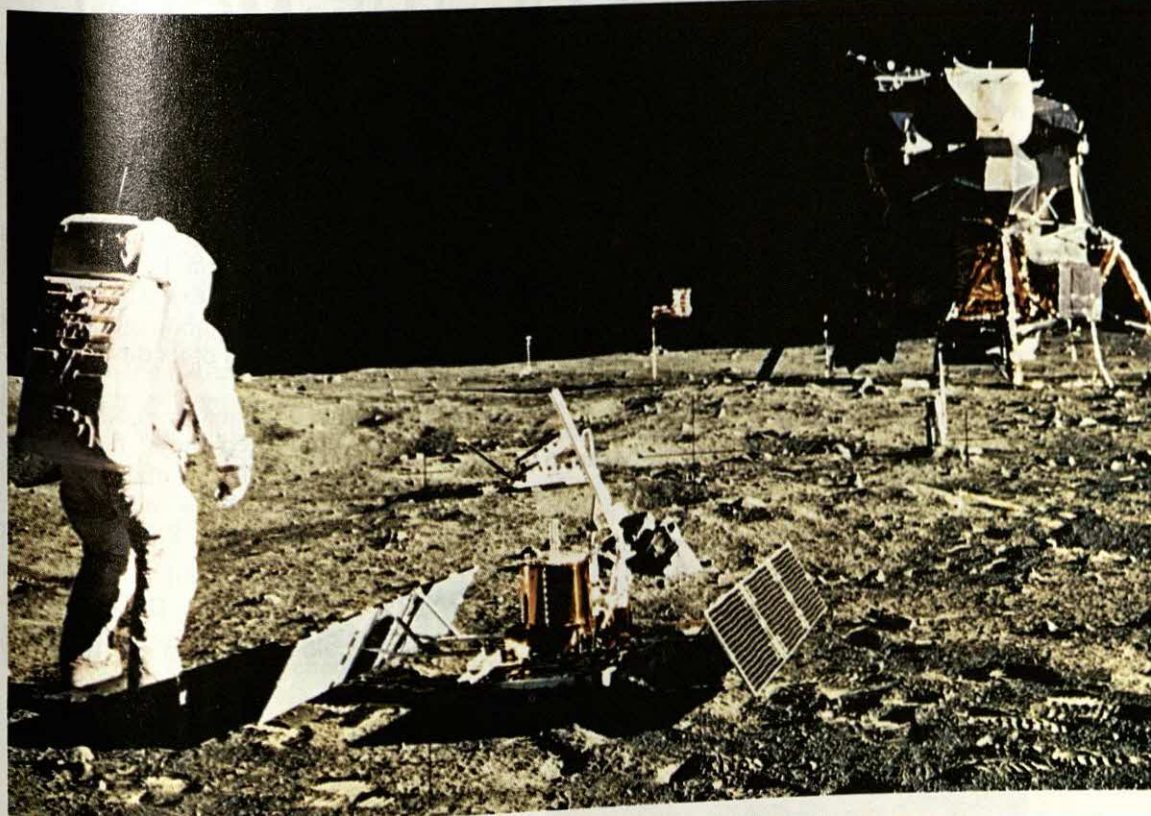
Astronaut John Young driving the Lunar Rover. The rover moved about the moon's surface at an average speed of about 6.5 kilometers an hour.





Left: Astronaut Eugene Cernan climbing aboard the Lunar Rover. Below: Tranquility Base, the landing site of the first Apollo mission. The lunar horizon is about 3 kilometers away.

NASA





NASA

A microsection of one of the rocks brought back to earth from the surface of the moon. This colorful igneous rock contains three minerals: clinopyroxene, plagioclase, and ilmenite.

GEOLOGY OF THE MOON

by Paul D. Lowman, Jr.

When man first turned the telescope on the moon over three centuries ago, he saw mountains and plains and craters. He realized that here was a world not unlike the earth in many ways. Astronomers mapped the side of the moon that faces us, naming its features and measuring their heights and depths. Soon they began to speculate on the origins of these features.

As the study of the earth—geology—developed along with astronomy, the nature of the moon's surface forms began to take on more significance. Scientists began to compare the earth and the moon, finding not only striking similarities but also great differences between them.

With telescopes and even more sophisticated instruments such as radar, astron-

omers could go only so far in understanding lunar features. In the 1960's, much greater knowledge became available when unmanned spacecraft photographed the moon at close range, landed on it, sampled its rocks, and sent the data to earth. In 1969, men first set foot on the moon and returned to earth with specimens of its dusty surface and rock formations.

All the data obtained from earth-based observation, unmanned moon flights, and the explorations of the Apollo astronauts have been combined into a single scientific field that has come to be known as *lunar geology*.

LUNAR AND TERRESTRIAL GEOLOGY

Lunar geology is essentially the study

of the moon through its rocks and landforms. Its goal is a correct and reasonably detailed picture of how the moon has evolved. It may be surprising to hear that we may soon be able to understand the moon's geologic development more clearly than that of the earth. This is especially true of the early stages—the first 2,000,000,000 years or so—of the moon's history. On earth, rocks representing this particular span of time have generally been deeply buried by younger rocks.

Even where such very ancient earth rocks have been exposed, they have often been metamorphosed, or greatly altered, by different geologic forces and events, almost beyond any resemblance to their original condition. In fact, they have frequently been destroyed by erosion or melting. Their former presence in many cases can only be detected by the gaps left in the sequence of the remaining rocks. The moon apparently has suffered far fewer such changes. By no means geologically “dead,” our satellite, compared with the earth, is in something of a geological “deep freeze.”

SURFACE OF THE MOON

Although the moon may at first seem to be a chaotic wasteland, its visible features may be sorted and classified. The surface is divided mainly between dark, generally level plains—the so-called *maria* (singular: *mare*)—and bright rugged *uplands*, or *highlands*. Covering both the maria and uplands are circular structures called *craters*.

Sometimes running for hundreds of kilometers across maria and uplands alike are long cracks, grooves, or valleys, called *rills*, which may be straight or crooked. Other kinds of features include ridgelike *wrinkles* in the maria, and probable volcanic forms such as *domes*, *cones* and apparent *lava flows*. We will take up these lunar features in turn, starting with the most conspicuous and puzzling, the craters.

Craters. The most striking lunar landforms are the craters: roughly circular structures, each consisting of a wall or rim that surrounds a generally somewhat level area. This inner area, or crater floor, may be depressed below the general level of the

ground surrounding the crater. The floor may be relatively smooth or may have mountains or hills. Craters may appear “fresh,” with raised, strongly defined, or rugged rims. Others are very shallow and virtually rimless. The latter type is apparently badly eroded, buried by later rocks or loose deposits or filled with solidified lava (once-molten rock).

Good examples of large craters are Copernicus and its very similar cousin, Tycho, to the south. Both craters are broad, rather shallow depressions, with raised edges and surrounded by piles of material that have evidently been ejected, or thrown out, from the craters by some means. This ejected material, or *ejecta*, forms a bumpy or hummocky terrain near the craters, and grades outward into hills that form a radial pattern.

These radial hills in turn grade into a halo of so-called secondary craters. The latter are believed to be shallow pits or gouges dug into the lunar surface by falling fragments of rock—the *ejecta*—thrown off from the main craters. Radiating still farther from the main craters are the *rays*: bright streaks along the lunar ground. Along these rays, chains of secondary craters are visible in many areas. The rays of crater Tycho extend almost halfway across the near side of the moon.

Inside the craters Tycho and Copernicus are concentric terraces going down along the walls toward the central depression. These terraces are probably blocks of rock that have slumped or slid down at some time after the craters were formed. Some sort of material also seems to have flowed down the crater sides toward the floor in many places.

In 1969, man first set foot on another member of the solar system. He returned with priceless treasures: rocks, soil, glass beads. Since that time, other astronauts have viewed earth from the lunar surface. The photo below shows the earth as a crescent above the moon's horizon.

NASA





NASA

Close-up of a moon rock. One specimen brought back by the astronauts is dated as 4,600,000,000 years old—the calculated age of our solar system.

The crater floors themselves may have been liquid at one time; that is, the craters were partially filled with lava, or molten rock. This seems evident from the ropy nature of the floor's topography, as in Tycho. Copernicus probably had a similar floor at one time. But since this crater is somewhat older than Tycho, its floor has been partly eroded. On the floors of these and many other craters, there are rises called *central peaks*, as well as many smaller hills.

Copernicus and Tycho are typical of certain of the moon's large craters. Many more are obviously very old, eroded versions of these two. Eratosthenes, for example, just east of Copernicus, once resembled the latter crater. But with the passage of time, small meteorite impacts and other processes have subdued Eratosthenes' topography; its rays are now invisible. Other craters resemble Tycho and Copernicus in most respects, except that they have been filled with dark rock of some kind, probably deposits of lava.

An extremely important question now arises: How big can craters be? This depends on the definition of a crater. Tiny craterlike pits of microscopic size have been detected in small fragments of moon rock. And yet we can consider many of the dark plains, or maria, of the moon as craters of a sort, because they are often surrounded or partly surrounded by wall-like ramparts of mountains.

If such a wide variety of landforms can be labeled "craters," we may well ask

whether they were all formed by a single kind of geologic process.

Origin of lunar craters. The origin of the lunar craters has been debated for centuries. Present-day thought centers on two main theories of crater genesis: (1) meteoritic or cometary impact; or (2) volcanic activity.

The impact theory has rapidly gained favor recently, especially among scientists in North America since the 1950's. Evidence supporting it comes from a variety of sources. First, sizable bodies—meteorites or possibly even very small asteroids—have been known to strike the earth occasionally. Many smaller bodies, however, are destroyed by the earth's atmosphere before they reach the ground. The moon has no atmosphere to speak of, so it would suffer many more meteor impacts for a given time in any given area than the earth would.

Second, there are many large, apparently impact craters on the earth. Meteor Crater, one and one-half kilometers wide, in Arizona and the Manicouagan structure, 60 kilometers across, in Canada are probable examples. Nearly fifty of these formations show evidence of what seems to be meteoritic impact.

This evidence consists of fragments of iron-nickel, a common component of meteorites; melting and recrystallization of rock caused by the impact shock of a meteorite landing; and unusual minerals that form under very high pressures and temperatures, which are usually produced by impact.

Third, many lunar craters, including Copernicus and Tycho, resemble earth craters that have resulted from the impacts of rockets and from explosions. The proportion of depth to diameter in many lunar craters equals that of earthly explosion craters.

Fourth, asteroids closely approach the earth and moon occasionally. In the past few thousand million years, at least a few such bodies must have hit the moon. If Copernicus and similar craters are not the scars of such encounters, where are the scars?

Advocates of the volcanic theory of

crater formation discount the arguments given above. They point out that the apparent impact origin of many terrestrial and lunar craters has not really been proved. Many lunar craters resemble certain huge earth craters, named *calderas*. A caldera is formed by the walls of a volcanic crater collapsing, enlarging the latter. The biggest known calderas, however, are nowhere near the size of the larger lunar craters.

Nevertheless, the volcanic theory of the genesis of such craters as Copernicus or Tycho and others like them appears at present to be the minority opinion among scientists. Despite this, however, there are moon craters that even impact theorists will admit are volcanic in origin.

The photos below show thin sections of lunar rock. The colors in the sections are caused by the interaction of polarized light with the crystalline structure of the various minerals. Each color appearing in the sections thus exposed usually represents a different mineral.

NASA



These volcanic craters tend to be much smaller generally than the average impact craters. Most of the chain craters—that is, those that tend to occur in lines or rows—are almost certainly volcanic. They are thought to arise, like many earth volcanoes, along cracks in the crust, through which molten rock or gas wells up, or erupts.

Some of the small craters may have been formed by the impact of rocks falling to the ground after they had been shot out in volcanic eruptions. Such volcanically caused impact craters would be called secondary craters, like the secondary craters around large impact craters.

Craters bear some resemblance to the circular maria: the latter also tend to have a circular shape and are often surrounded by rugged land not unlike the crater walls in appearance. Maria, like many craters, have dark floors of solidified lava.

Maria. In contrast with craters, maria are generally far larger and cover thousands of hectares of the moon's surface that faces the earth. Like many lunar craters, maria probably originated in impacts of huge bodies from space landing on the moon. A *mare basin* was excavated by a falling body. Later this basin became filled with lava flows that then hardened to form the mare itself.

Maria are generally level, but often have low ridges and domelike elevations. Tremendous numbers of smaller craters pock the marial surfaces. The ridges resemble wrinkles and may either be folds in the rock or cracks through which lavas have flowed.

The domes may also be volcanic. (The name *dome* is also given to hills in the upland regions.) The Marius Hills are classic domes. They seem to be volcanoes representing late activity following the major eruptions that produced the rock of the mare regions.

The maria are generally dark, in contrast to the surrounding highlands, which are brighter in appearance, more rugged, and stand higher than maria.

Lunar highlands. The lunar highlands, or uplands, make up much of the moon's earthward side and almost all of its far side.

The reason for this strange topographical situation is not certainly known at present (although some theories have been advanced). The highlands are rough and densely pitted, with craters formed in or on other craters.

The greater number of craters in the highlands than in the maria seems to indicate that the highlands have been bombarded by meteorites or similar bodies a much longer time. Therefore, as landforms the highlands are probably older than the maria. The original lunar crust, if it still exists or is exposed anywhere, must be part of the highlands.

Another type of lunar surface feature that adds to the ruggedness of the moon's face is the aforementioned rills, which may cross highlands and maria. They resemble cracks or valleys in the moon's crust.

Rills. Rills have been the subject of speculation since the first telescope surveyed the moon. Some are straight and look like gigantic breaks, or faults. Huge blocks of the lunar crust probably have slipped down along these faults on one side or the other.

A different sort of rill meanders wildly for many kilometers, like the twists and turns of a river. This appearance in fact suggested that these rills were once river valleys at a long-past time when the moon supposedly had an atmosphere and water. However, none of the lunar rock samples analyzed thus far show more than the slightest trace of water, if any at all. According to all currently available evidence,

NASA



Dr. Harrison H. Schmitt, a civilian geologist, made valuable observations of the lunar landscape during Apollo 17's lunar exploration. Here, Schmitt is shown collecting rock samples.



NASA

Photomicrograph of a thin slice of moon rock, taken in polarized light, reveals mineral structures in a beautiful array.

our moon never had any appreciable amount of surface water.

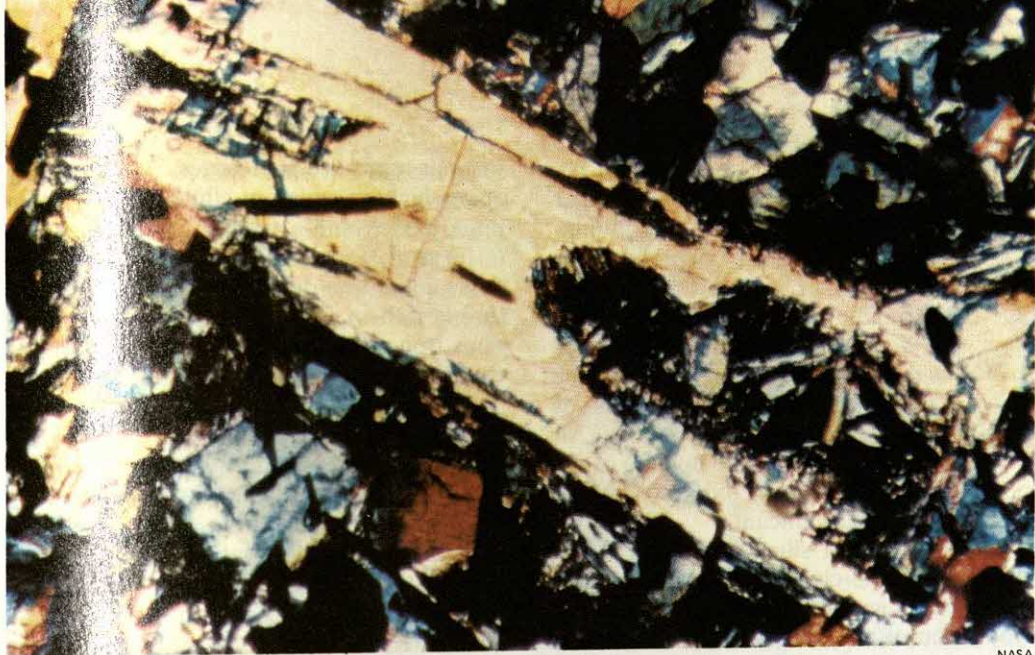
Another possibility is that the sinuous rills are channels for the drainage of lava from volcanic activity. Still another is that they are valleys dug by the flow of volcanic ash, which is a fine-grained material shot out in some volcanic eruptions.

In this survey of the moon's face, we have mentioned from time to time the rocks and other material of which its features are composed. Now that samples of the lunar surface have been brought back to earth, scientists have been able to determine accurately the nature of our satellite's rocks and the probable composition and structure of the moon.

LUNAR ROCKS AND MINERALS

Almost everywhere, the surface of the moon is covered by layers of dust and larger rock fragments. The exact thickness of these layers is not known everywhere for certain, but probably averages at least several meters. In some areas it may be much deeper. This so-called lunar dust, "soil" or mantle is technically known as the *regolith*.

Regolith. The regolith was probably formed by several processes. These would include the impact of falling bodies break-



NASA

ing up the solid rock; volcanic eruptions scattering debris over wide areas; landslides; moonquakes; and the pulverizing effect on rock of intense radiation from the sun and from outer space.

The lunar regolith contains a variety of materials: unaltered fragments of the underlying bedrock; bits of melted rock and glasslike globules; volcanic ash; and small amounts of meteoritic matter. The glasslike substances, especially the spherules, were probably produced by the impact of meteorites. The latter, by their fall, shattered and melted the moon rock, which solidified again, but into glasslike material.

Much of the regolith consists of so-called *breccia*—a deposit of sharply angular rock fragments of large size compacted together with various other mineral pieces. The regolith generally has a loose consistency. Here and there the broken material, a mixture of native rock and meteoritic material, is cemented together to form a kind of compound, secondary rock.

Some samples of regolith, however, are unlike the average. They contain a high proportion of potassium, phosphorus and unusual metals known as rare-earth elements. Scientists believe that this portion of the regolith represents the remains of an ancient crust that once covered the moon's surface about 3,500,000,000 years ago. The crust was destroyed later by volcanic eruptions and meteorite impacts.

In this sample, the yellow and reddish-orange mineral is pyroxene, the black is ilmenite, and the whitish-blue is plagioclase (a type of feldspar).

The regolith covers the bedrock of the moon. The bedrock's nature has been inferred from the larger rock pieces, which are fragments of this bedrock.

Solid lunar rocks. The moon's solid rocks and a certain proportion of its regolith seem to be of *igneous* origin. *Igneous rocks* are rocks that form directly by the solidification of molten rock matter, or *magma*. Scientists have determined that most lunar rocks are similar to earthly igneous rocks called *basalts* and *gabbros*. Basalts and gabbros are heavy and generally dark in color.

Lunar rocks are also generally heavy and dark. They range from fine-grained to coarse-grained material. Basalt is very finely crystalline, whereas gabbro is relatively coarse. The grains may be of equal size or varied. The lunar specimens obtained thus far are generally vesicular—that is, full of cavities. It is believed that these cavities were produced by gas that was trapped for a while in the rock when the latter was molten. The gases then escaped into space when the liquid rock poured out onto the lunar surface. Basalts and gabbros are rich in minerals containing a high proportion of metals such as iron, magnesium, aluminum, calcium and titanium.

Lunar minerals. Elements seldom occur native, or free, in earthly and lunar rocks. They are often combined into mixtures of natural chemical compounds, or minerals. The most common types of mineral compounds are *silicates*—combinations of silicon and oxygen on one hand with various metallic elements on the other—and the *oxides*—combinations of oxygen with various other elements. An important oxide is silica, a compound of silicon and oxygen. Various oxides of iron or titanium are also significant.

The variety of silicate minerals is tremendous. Their classification depends on the kinds of metals they contain and the proportion of silicon and oxygen. The silicates of importance to the subject of this article are the feldspars: silicates of potassium, sodium, aluminum and calcium; and the pyroxenes and olivines: silicates of calcium, magnesium and iron. The olivines are lower in silicon-oxygen than the pyroxenes.

Most of the minerals found in moon rocks are much like their earthly counterparts. Several, however, are completely unknown on earth. The important minerals

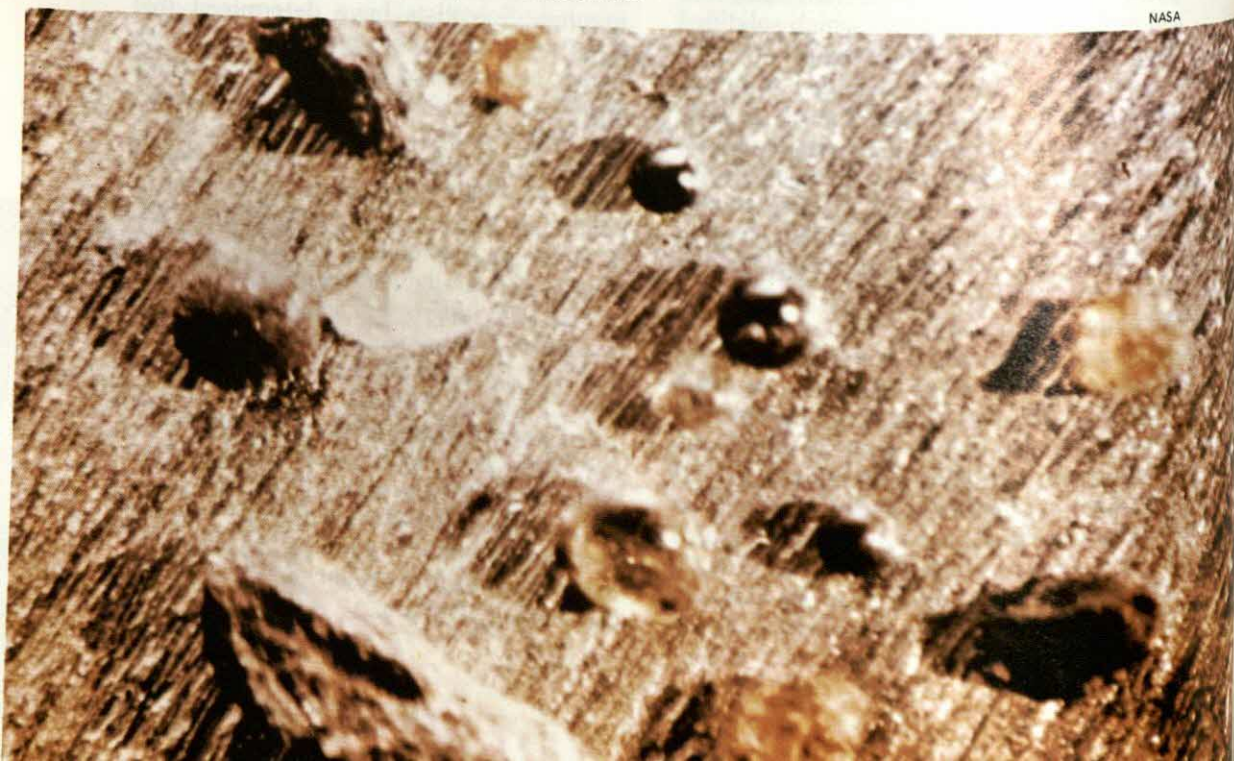
in lunar rocks are pyroxenes, olivines, feldspars high in calcium, ilmenite (iron-titanium oxide), and a volcanic form of quartz (a form of silica) called *crystalite*. All of these and others point to a lunar chemistry that is rich in calcium, magnesium, aluminum, iron, titanium and chromium, more so than on earth. In contrast with the earth, however, moon rocks are relatively lower in potassium, sodium, silicon, oxygen and water.

There seems to be a scarcity of any granitic rocks in the moon. Earthly granites are rich in minerals that contain much sodium, potassium, lithium and silica. In contrast with basalts and gabbros, granites are less dense and are lighter in color. On the earth, granites make up much of the continents, while basalts mainly make up the crust of the earth underlying the different ocean basins.

Lunar rock history. From the chemical and physical nature of the lunar rocks, scientists have deduced that many of them cooled quickly. That is, they were lavas that erupted onto the moon's surface, "froze" rapidly and produced the very fine-

A microscopic view of glass spherules found in lunar soil. Scientists believe these beads were formed under very high temperatures, perhaps as a result of the impact of a meteorite.

NASA



grained rock called basalt. Magmas that cooled more slowly produced coarser-grained rock—gabbro. Gabbro is usually found in igneous rocks that have been intruded, or pushed, into surrounding older rocks, not poured out onto the surface of the ground.

Geologists have also gained insight into the kind of magma that the lunar rocks came from. It was probably very fluid, or "runny," and may have had a temperature as high as 1,300° Celsius. It was also a "dry" magma, or low in water. This magma may form—or once formed—a reservoir of liquid inside the moon or may have been produced by local melting of lunar rocks at some depth underneath the surface. Lunar rocks also indicate something about the moon's past and present environment, such as the presence of an ancient magnetic field and the intensity of cosmic rays in the past.

Lunar environment. The almost non-existent water in most rock samples seems to show that the moon never had any appreciable water supply, especially at the surface. The fact that free iron has been discovered in a number of rock specimens shows that atmospheric oxygen and water have scarcely been present on the moon. Iron is speedily attacked by any water and oxygen to form hematite (rust) and other minerals.

A number of lunar rocks show evidence of shock or impact. They have been struck repeatedly by bodies, probably from space, so that parts of the moon rock melted and then solidified into glasslike matter.

Some of the solid-rock and regolith specimens were found to contain gases of various kinds. These gases probably came from two chief sources: radioactive decay of certain mineral elements, such as uranium and thorium; and radiation particles from outer space. In the latter case, the particles penetrated the lunar surface and became trapped there. In effect, these particles are bits of matter and energy from the sun and stars and not parts of the moon.

Densities and distributions of moon rocks. Some of the maria each have a mass concentration ("mascon") of dense matter under the surface. Mascons were detected by the fact that vehicles orbiting the moon



NASA

A photomicrograph of a thin section of lunar rock, for use in mineral identification.

are pulled slightly down toward those maria. Scientists are still unable to explain what mascons are, although they have some theories. Under a number of craters, on the other hand, the rock densities are less than in the surrounding land. Vehicles flying over these craters rise slightly higher in their orbits.

Astronauts have explored lunar mountains and maria. They have found highland rocks lying in the maria. The rocks were thrown there by the impact of huge meteorites landing in the highlands. These rock specimens are quite unlike the mare rocks in composition and, in fact, markedly resemble certain strange rocks found on earth in different areas.

These unusual earth rocks, and their lunar counterparts, are called anorthosites, because they consist almost entirely of a calcium-rich feldspar called anorthite. Other minerals and their elements are present in smaller amounts, most notably ilmenite (titanium-iron oxide) and minerals containing chromium and aluminum. Anorthosite is chemically related to gabbro. On earth, there are a variety of anorthosites widely scattered in the crust.

All known earthly anorthosites are very old. The oldest known at present is well over 3,000,000,000 years in age, which is comparable with the ages of moon rocks. Anorthosites are less dense than basalts and gabbros. The parent magma of anorthosite must have been very hot and fluid.

These facts suggest some kind of similarity between the moon's history and the



Undisturbed lunar sediment, photographed as core tube is opened. The material resembles silty sand, with scattered rock fragments and glassy particles.

NASA

Examining a moon rock at the Manned Spacecraft Center's Lunar Receiving Laboratory in Houston, Texas. The structure and composition of each rock sample was carefully analyzed.

NASA



earth's early development. Is it possible that the earth's original crust was anorthositic? Did anorthosite at one time form the continental masses of the earth instead of granite? If so, the anorthosites on earth have largely disappeared, being replaced or swallowed up by later outpourings of basalt and granite.

Rocks in the lunar highlands are different from the rocks of the maria. They contain far more minerals and are also unlike the lowland material in composition. The structure of the rocks in the mountains is very complex. They consist of old breccias composed of still earlier clumps of rock fragments. Evidently they underwent a complicated development. Whether the rocks gathered by Apollo astronauts in the highlands are typical of all rock in the lunar mountains remains to be seen.

AGE OF THE MOON

In general, the surface features of the moon and its rocks are very old, on the order of several thousand million years. How such figures for dates of rocks and geologic formations are derived is very interesting.

First, geologists are concerned with the *relative* ages of features; that is, which is older or younger than which. In general, the older a rock or landform, the more worn or eroded it tends to be. It may also be chemically and physically altered. Younger rocks and features tend to overlie the older ones, but this is not always the case. For example, a mass of molten rock may shove its way up from below into older rocks and so come to lie beneath them. The study of the sequences of rock layers, or strata, as well as of other kinds of rock masses, is known as *stratigraphy*.

Second, geologists want to determine the *absolute* age of a rock, in terms of years. This is usually done by chemical and radioactivity measurements of the types of elements in the rock.

The amount of time that a rock, regardless of its exact age, has been exposed to radiation, especially from the sun or from outer space, can also be determined from its chemical composition and radioactivity. In this way, scientists have found out how

long meteorites and surface moon rocks have been exposed to the conditions of space. This is a valuable clue to their history.

Dating lunar rocks. By earth standards, lunar rocks are very old. The youngest known rocks of the moon are well over 3,000,000,000 years in age. Other rocks are more than 4,000,000,000 years old. The lavas of Mare Tranquillitatis crystallized about 3,700,000,000 years ago; those of Oceanus Procellarum, a few hundred million years later. Mare Imbrium was formed around 3,900,000,000 years back and is thus one of the oldest lunar features.

Strange to say, the small fragments of the regolith proper are actually *older* than the supposed underlying solid rock of the maria explored thus far. They range up to 4,600,000,000 years in age. Here then is a kind of "reversed" stratigraphic sequence where overlying deposits are older than the lower ones. The regolith at the Apollo sites has been lying on the lunar surface, exposed to solar and space radiation, for periods of a million to several hundred million years.

This stratigraphic problem of the lunar crust is a puzzling one that has led geologists to consider a number of new theories in order to explain it.

Lunar stratigraphy. To summarize the situation on the moon's surface, we have the presumed solid rock of the moon's crust at some depth below the regolith. At least part of the latter must have been derived from the bedrock, by meteorite impact, radiation and other erosional forces.

To explain why the regolith generally is older than the bedrock, geologists have considered several alternative theories. One is that the regolith at any point on the surface is a representative mixture of pieces of the lunar crust as a whole. Fragments from wide areas of the moon, including the very ancient highlands, may have been scattered all over for vast distances by large and continual meteorite impacts and volcanic eruptions. Thus the age of the regolith at any one place may represent the age of the lunar crust as a whole. Another possibility is that the bedrock was pushed as a magma up into the already existing

regolith, which is thus older than the bedrock.

It is obvious that the age of the moon's rocks as a whole is very great, and the moon as a body is probably older still—perhaps 5,000,000,000 years. The oldest rocks discovered so far on earth are about 3,500,000,000 years or so in age. This does not necessarily mean that the earth is younger than the moon: simply perhaps that the earth's oldest rocks have not yet been found, have not survived, or have been altered beyond accurate measurement.

Let us now combine these various jigsaw puzzle pieces of knowledge—topography, rocks and ages—into some kind of comprehensive picture of the moon's probable evolutionary history.

EVOLUTION OF THE MOON

Formation of the moon. The moon's development of course began with its formation, about which even at present there is little agreement among scientists. The chief theories hold that the moon was born from the earth at some early or perhaps relatively late stage in the latter's history; or that the moon and earth developed independently from generally the same materials at the time the entire solar system originated; or that the moon was captured by the earth when the former was already in existence. In these theories the earth and the moon may be of the same age or either one younger than the other.

A theory popular at one time said that the moon split off from the earth, which was once more massive than now. George Darwin, a nineteenth-century scientist, speculated that this occurred when the early earth was molten. The idea has been revived recently by some authorities.

A widely favored theory today is that the moon formed from the same cloud of gas and dust as the earth. This happened at the same time or probably later than the formation of the earth. Despite the fact that so much iron occurs in the surface rocks of the moon, the overall density of the moon is low compared with that of the earth—about 60 per cent of the latter's. In other words,

much of the moon below its surface must be considerably less dense than the crustal rocks. This may be accounted for by considering the moon as a whole to be deficient in iron. The earth, on the other hand, is thought to have a heavy metallic core at its center that is very high in iron. This situation accounts for the earth's greater density, although many of its crustal rocks are not so rich in iron as the corresponding moon rocks.

To explain this theory, a number of scientists believe that the moon formed *after* the earth's iron core developed. One way in which this could have happened is that heated silicate minerals evaporated from the intensely hot early earth. These minerals actually formed clouds or rings of sediment, or rock matter, hanging in the thick atmosphere of our world at that period. The moon formed by the condensation and accumulation of these clouds.

Early growth of the moon. During or shortly after the moon's genesis, by whatever process, the moon probably became very hot. This condition was caused by the compression of materials forming the moon, by heat produced in the radioactive decay

Astronaut Edwin Aldrin taking a sample of regolith. The Apollo 11 mission collected a quantity of lunar rocks and soil for study. This was the first real piece of lunar geological research.

NASA



of short-lived radioactive elements and by heat from other sources. The high temperatures may have caused extensive melting and remelting of lunar rocks, enough so that different kinds of igneous rock were able to develop. This may account for the possible difference between basaltic plains (maria) and anorthositic, or possibly even granitic, highlands.

While all this was going on, a great number of bodies called *planetesimals* from space fell into the moon, contributing their share of elements and forming the oldest or at least very old craters, in the highland areas and on the far side. The infall of these bodies was the last stage of the moon's first growth period.

Second stage of moon growth. During this time, several bodies, something like asteroids or small moons, were growing in the neighborhood of the moon. These were swept up by the moon, which thus suffered catastrophic impacts, impacts that formed the basins later occupied by the circular maria: Serenitatis, Crisium, Humorum, Imbrium, Orientalis, and others. These tremendous impacts caused some melting of the lunar rocks, but did not really produce the large-scale lava outpourings that actually filled the basins and so formed the maria afterward. Many impact craters, such as Archimedes and Plato, were formed after the "excavation" of the mare basins, but before these basins filled up.

Still more geologic features arose before the mare lava flows. Volcanic activity in the lunar highlands (or in what are now the highlands) filled craters such as Ptolemaeus and Alphonsus with lava and ash. This material resembles the later mare lavas and ash, but is more cratered because of its greater age. The highland lavas are also more glassy because of the larger number of impacts causing remelting that they have undergone.

We now come to an event in lunar history that has actually been dated by man: the lava eruptions that produced Mare Tranquillitatis, where the landing site of Apollo 11 is located. This igneous activity took place 3,700,000,000 years ago, which is the time since the crystallization of

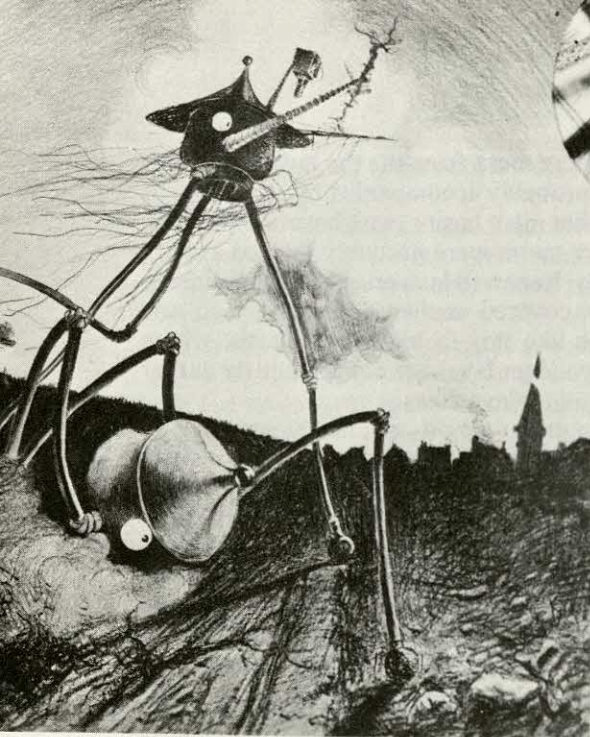
the mare rocks from the the lava there. This was probably accompanied by similar flows in other mare basins, which were thus filled. Other maria were probably formed at later times. Renewed lava eruptions from time to time covered earlier flows. We can see areas like this in Mare Serenitatis, where the younger lavas are conspicuously darker than older lava flows.

Latest period of lunar history. By about 3,000,000,000 years ago, the moon's face attained essentially its modern appearance. However, lunar geologic evolution continued in the form of occasional local volcanic activity in the maria and the lunar highlands. The volcanoes of the Marius Hills are from this relatively late period, as well as the chain craters, sinuous rills and darker highland regions.

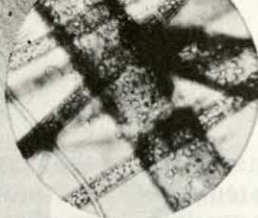
The most spectacular events of this time, however, were the impacts that produced the giant craters, such as Copernicus. The agents were probably huge meteorites or comets moving at terrific speeds. As such a body smashed into the lunar surface, it threw up a huge cloud of debris, from which long streamers shot out in all directions for hundreds of kilometers. As the streamer materials settled on the surface, they formed the ray systems of craters such as Copernicus, Aristarchus, Tycho and others. The strange red spots that have been seen in Aristarchus may be volcanic activity still lingering after the initial impact so many years ago.

In the meantime, a slow, uneven rain of minor meteoritic material keeps falling onto the moon. Their impacts have been recorded by the seismometer (quake detector) left in Oceanus Procellarum as occasional "events." The result of this rain has been the countless small and even microscopic craters that riddle the moon's rocks and mantle. The mantle, or regolith, of course, itself is the result of this process in large part.

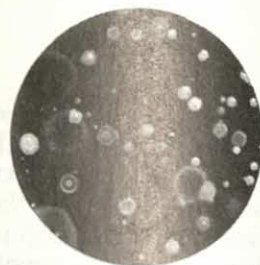
Much has been learned about the moon in only a few years. But discoveries bring new mysteries. That scarred globe, the moon, which has circled earth for so many ages, will be the target of investigation for years to come.



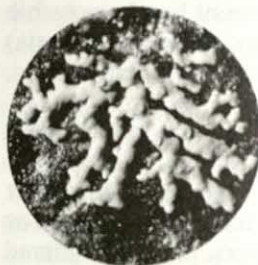
The Bettman Archive



ALGAE



BACTERIA



LICHEN

Robert W. Lichtwardt
Brooklyn Botanic Garden
Rutherford Plant

Left: drawing of Martian invaders in H.G. Wells' novel "The War of the Worlds." Above: earthly algae, bacteria, and lichen—simple forms of life that may resemble possible life forms on Mars.

LIFE ON OTHER WORLDS

by Richard S. Young

"Two large dark-colored eyes were regarding me steadfastly. The mass that framed them, the head of the thing, was rounded, and had, one might say, a face. There was a mouth under the eyes, the lipless brim of which quivered and panted, and dropped saliva."

Thus did the noted English writer H. G. Wells describe a Martian in his famous novel *The War of the Worlds*, published in 1898. He told how the Martian invaders of the earth, octopuslike monsters with an intelligence far beyond man's, nearly destroyed human civilization.

Wells' story, however fanciful, still reflects the thinking of many people about possible living beings on other worlds. Man has long regarded the stars, sun, moon and planets as the homes of gods and demons. In countless myths, man is either the victim of these creatures or is helped and civilized by them. In some tales, the extraterrestrial (nonearthly) beings are fully human, at least physically. These themes have survived in

present-day science fiction and fantastic stories, including many flying-saucer reports. Many people believe that well-developed life exists somewhere in the universe besides here on earth. They believe that some of these possible living forms may be as intelligent as man, if not more so.

At the opposite extreme, many people, including a number of scientists, look upon worlds beyond the earth as lifeless. Or, at best, these worlds are thought to have only the lowest forms of life, such as bacteria, algae and lichens. These ideas are more correct than the others, at least in regard to our solar system.

The branch of biology that deals with the possibility of life on other worlds is called *exobiology*. It literally means "biology outside [of the earth]." Exobiology deals in addition with the fundamental question of the origin of life. In this article, we examine the prospects for life beyond the earth and describe some of man's attempts to communicate with extraterrestrial civiliza-

tions. First, however, let us consider the chemical foundations of life and how it may arise on a planet.

THE CHEMISTRY OF LIFE

The only life we know of at present is the kind that exists on the earth. This life is extremely varied and includes bacteria and protozoa, which are so small that they can rarely be seen without a microscope; millions of species of insects; giant sequoia trees, and whales; and, finally, man himself. The bodies of all these organisms are composed of chemical elements. Relatively few elements are found in living matter: carbon, hydrogen, oxygen, nitrogen, sulfur, phosphorus and some others, including certain metals. These elements are usually combined into biological compounds, or biochemicals. A number of biochemicals, such as proteins, contain all or most of the elements listed above. Other life-related molecules, such as water and carbon dioxide, contain only a few of these elements. Most biochemicals are organic compounds; that is, they contain carbon. Because so many biochemicals contain carbon, life on earth is said to be carbon-based.

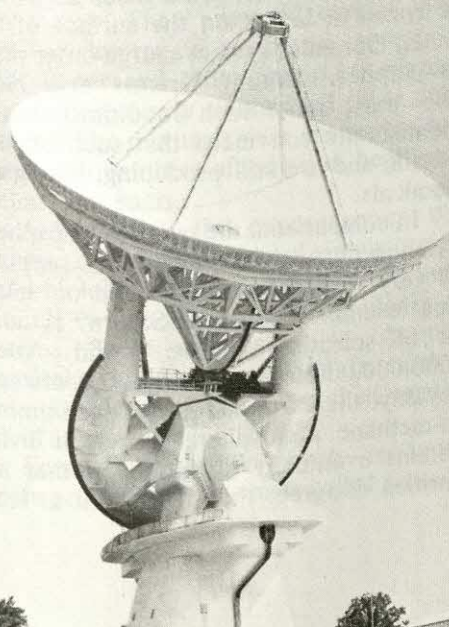
There are three major reasons why exobiologists usually limit their work to carbon-based biochemical systems. First,

these systems represent the only forms of life with which we are familiar. Second, astronomers have detected simple, earth-type organic compounds in outer space. This discovery has led many exobiologists to conclude that the chemical building blocks of life are the same everywhere in the universe. Third, these chemicals become part of planets when planets first form from the elements in a nebula, or cloud of dust and gas in space. The elements are present in proportion to their abundances in space: hydrogen, helium, carbon, oxygen, nitrogen and so on, with a huge excess of hydrogen. In the atmosphere of a primitive planet, compounds such as methane, water and ammonia have been formed from many of the elements.

MATTER IN OUTER SPACE

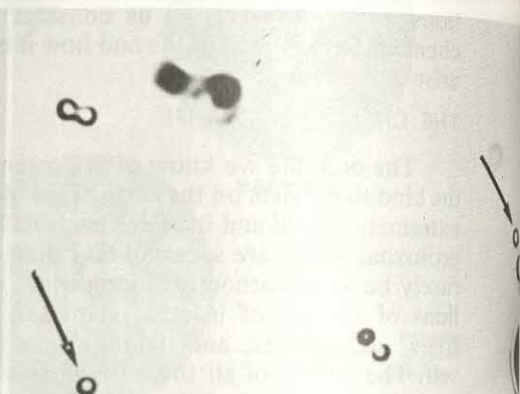
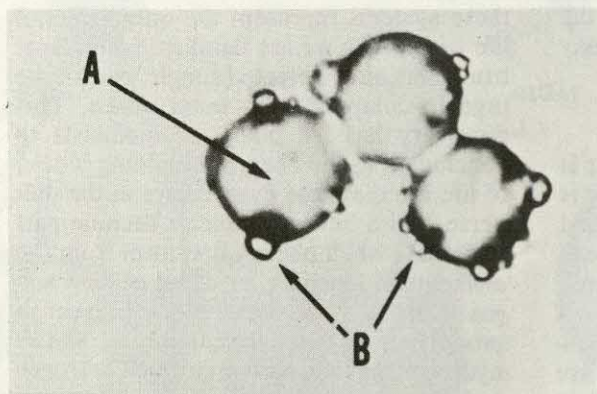
Outer space contains a variety of matter, ranging in size from subatomic particles to cosmic dust and larger bodies. But by earth standards, outer space is a vacuum. Its average density of matter is very low, perhaps as low as one molecule per cubic centimeter. In comparison, the atmosphere of the earth is millions of times denser.

Like all matter, cosmic matter exists in various states of energy. Therefore it emits different wavelengths of electromagnetic



A large radio telescope: this type of instrument receives natural radio waves emitted by cosmic matter and other bodies in outer space. Radio telescopes are also being used to try to detect possible electronic messages or signals from advanced civilizations that could exist elsewhere in the universe.

National Radio Astronomy Observatory



all photos this page and next, Institute of Molecular Evolution, University of Miami

radiation. If the electromagnetic waves are in the radio range of frequencies, they can be received by radio telescopes on earth. Scientists analyze these waves to identify the chemical nature of the atoms and molecules that emitted them.

Among these cosmic molecules, scientists have already identified a number of chemicals that are related to life on earth. That is, they are identical to elements and compounds that are part of living systems, that help life to exist or that are produced by organisms. The substances discovered in space include water, ammonia, methane, formaldehyde, cyanogen, hydrogen and methyl alcohol. These compounds, except for ammonia and water (and excluding the element hydrogen), all contain carbon.

How did these chemicals originate? Most scientists believe that these substances are of inorganic, or nonliving, origin. That is, the biological chemicals evolved from nonliving chemical systems. This theory is called the chemical evolution of life. It is closely related to the origin and development of planets and the origin of life, in that it seems to be the same kind of chemistry believed to have occurred in the early history of a planet, preceding the appearance of life there. The question in some scientists' minds is whether these biological compounds do lead to the formation of living things.

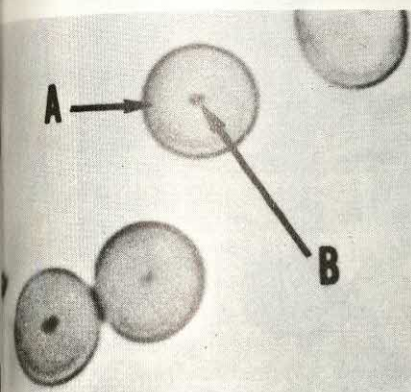
CHANGING CHEMISTRY OF PLANETS

Chemical evolution begins even before the appearance of stars and planets in a par-

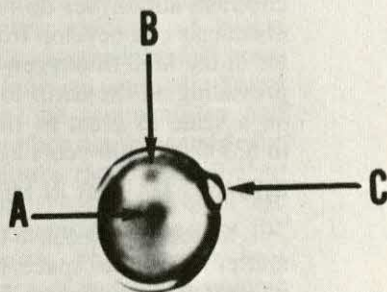
ticular part of space. Simple organic compounds are formed from elements in the cosmic matter. Slowly, cosmic matter of all kinds comes together to form stars and planets. As this happens, many of the organic compounds become part of the planets. Thus, at this early stage of development, the planets have a composition much like that of the space matter from which they formed. Later the chemistry of the planets changes as they develop through the ages.

Many scientists believe that a typical early planet, just after its formation, has an atmosphere with large amounts of hydrogen, methane, ammonia and possibly water vapor and oxygen. If it is distant from the star (sun) around which it revolves, the planet is cold. Therefore, much of its atmospheric gases may be condensed into clouds of droplets or snow; some gases may even be frozen in layers on the surface of the planet. Our solar system's large outer planets—Jupiter, Saturn, Uranus and Neptune—exist under such conditions. Many scientists are convinced that such planets contain, and are still producing, biological chemicals.

In comparison, the earth's atmosphere is rich in nitrogen and oxygen. It supports life very well, as we know. Did our earth once resemble Jupiter or Saturn? A number of scientists believe it did. About 4,000,000,000 to 5,000,000,000 years ago, they say, the earth had a hydrogen-ammonia-methane atmosphere in which living systems eventually developed. Other authorities disagree: they think that at least



Formation of proteinlike matter (proteinoids) into spheres (microspheres) superficially resembling living cells. Photos on preceding page: left, microspheres (A) forming buds (B); right, buds released from parent spheres by heat. Photos on this page: left, additional proteinoid matter (A) collects on buds (B), forming new spheres; right, new microspheres (B), produced from bud (A), puts out a new bud (C) in turn.



some free oxygen was also present in the earth's early atmosphere.

If planets start with atmospheres of hydrogen, methane and ammonia, does this always mean that biochemical compounds will form? Do these compounds always lead to the development of life on a planet? Experiments in the laboratory have given us some tentative answers to questions such as these.

CHEMICAL EXPERIMENTS

A Russian biochemist, Alexander I. Oparin, suggested in the 1920's that organic compounds could originate in a planetary atmosphere of hydrogen, methane and ammonia. This event could take place spontaneously if enough energy is present.

Beginning in the 1950's, scientists carried out experiments that confirmed Oparin's theory. They prepared various mixtures of hydrogen, methane, ammonia, water vapor and other gases. (A mixture did not necessarily contain all these substances.) The mixtures were then subjected to various quantities of radiation, heat, and/or electrical energy.

As a result, the experimenters obtained a surprising number of simple organic and biological compounds including amino acids, which are the building blocks of proteins. Some scientists also obtained chemicals that are contained in nucleic acids (DNA and RNA), the chief regulators of activities in living cells.

Some investigators even produced simple proteinlike molecules called *protein-*

oids. These proteinoids were organized into tiny globules, or *microspheres*. Microspheres resemble certain simple bacteria and carry out some functions of living cells. None of the experimenters claim that proteinoids and microspheres are alive. However, they think these structures could evolve into living matter (protoplasm) and cells, given enough time and the right environmental conditions.

In summary, visualize the earth, 4,000,000,000 to 5,000,000,000 years ago, as a planet shrouded in a cloudy atmosphere of ammonia, methane, hydrogen, water vapor and possibly other gases, including oxygen. As energy from the sun and perhaps from the earth itself acted on the atmosphere, more organic and biological compounds were produced.

Gradually these compounds fell from the atmosphere onto the earth's surface, where they formed lakes or ponds of thick, organic-chemical "soup." The synthesis of compounds probably continued here. Eventually, over a period of millions of years, the biochemicals developed into living cells.

More recently, some 3,000,000,000 to 4,000,000,000 years ago, the earth's atmosphere began to change; it became rich in nitrogen and oxygen. Some scientists relate this change to the rise of green plants, which released oxygen into the air. They also think that the present complex life of the earth could only have evolved in an oxygenated environment, although certain types of bacteria are known to exist in an

anerobic, or oxygen-lacking, condition. In contrast, authorities do not think that biochemicals can develop from inorganic matter in the kind of oxygen-rich environment prevailing on the earth today—at least not on a scale as great as that 4,000,000,000 to 5,000,000,000 years ago.

ORGANIC MATTER IN METEORITES

Meteorites, because they are pieces of matter from outer space, have long been of interest to exobiologists. Here, too, controversy has raged. Carbon-bearing chemicals, detected in a few meteorites, were usually dismissed as earthly contamination. Peculiar structures resembling earthly microorganisms such as bacteria and algae were also observed in some meteorites. Many scientists felt that these structures were entirely inorganic features. But other authorities accepted most of these findings as suggestive of nonearthly life.

The argument has been settled to some extent by scientists of the National Aeronautics and Space Administration (NASA) in the United States. They analyzed two meteorites with sensitive space-age instruments. Each of the stones—one that fell in Kentucky and the other in Australia—yielded 18 amino acids, plus other compounds of biological significance. Many of the amino acids are found in living systems on earth. But other meteoritic amino acids are unlike any found on earth. Thus the NASA scientists concluded that the biological compounds in the two meteorites originated somewhere else in the universe.

These compounds were probably not produced by living things. But they do indicate that the chemical evolution is indeed happening in different areas of the universe; biochemicals are forming from inorganic matter, and it is possible that these are leading to the creation of living things.

PLANETS FIT FOR LIFE

What are the chances that chemical evolution will lead to life on a planet? How many planets in the universe are suitable homes for living things? Astronomers estimate that there are 10^{20} (1 followed by 20 zeros) stars in the universe. Most of these



Scientist at the Ames Research Center displays a meteorite in which extraterrestrial amino acids, the building blocks of proteins, were found.

stars probably have planets circling them, just as our sun does. (So far, we have been unable to see any planets outside our solar system.) Astronomers estimate that life could have arisen on some 100,000,000,000 planets. Of these, many resemble the earth and could support the kind of life with which we are familiar.

We know that life is very adaptable. Nearly every place on earth is the home of some organism. Certain bacteria and algae survive temperatures close to the boiling point of water; others endure temperatures far below freezing. Therefore planets we think hostile to life may harbor living things of some kind, no matter how simple.

Discovering life of any sort on other worlds would do more than satisfy our curiosity. It would also shed light on the evolution of the earth and of the organisms that inhabit the earth. If planets and life systems are forming now in other parts of the universe, many of them may be at a stage that the earth passed through several thousands

of millions of years ago. Other planets and life systems may be older than the earth; these could show us what our future may be.

As we develop the ability and instruments to investigate other planets closely, we will obtain important clues to the earth's past and future. Even lifeless worlds may tell us something. If, for example, an other-wise earthlike planet is found to be barren of life, it would cast serious doubt on present theories of the chemical origin and evolution of life.

At present, we have direct knowledge only of planets in our own solar system. These have been studied closely for well over three centuries, since the invention of the telescope. The surfaces of Venus and Mars have also been investigated by means of unmanned space probes. This type of exploration is continuing and may be extended to other planets. Will scientists discover life on any of them?

MERCURY AND VENUS AS HABITATS

Mercury, the planet nearest the sun, is somewhat larger than our moon. Like the moon, it has little or no atmosphere. This fact, plus the very high temperatures, especially on the daylight side of Mercury, rules out the existence of appreciable water and organic molecules. As far as exobiologists are concerned, Mercury is not a high-priority object for study.

Venus, the second planet from the sun, is of greater interest. In size and density, it is very similar to the earth. But, because it is closer to the sun, it is hotter than our planet. Storywriters and some astronomers once pictured Venus as a tropical paradise of jungles and swamps, teeming with strange reptilian and amphibian creatures and blessed with eternal sunshine.

Radio telescopes and interplanetary probes have spoiled this rather attractive picture of Venus. From the exobiological point of view, Venus is really a grim place. The atmosphere is very unlike the earth's; it consists mostly of carbon dioxide gas, with small amounts of water and carbon monoxide. Dense clouds, whose exact composition still defies analysis, hide the planet's surface from sight.

The U.S. Pioneer Venus probe and the Soviet Union's Venera probes have provided close scientific readings of Venus (Venera 13, in 1982, returned the first color pictures of the Venusian surface). These data indicate that the temperatures and pressures at the surface are unbearable for humans. Temperatures range from 460° to 540° Celsius. Atmospheric pressures are about 100 times greater than those at the surface of the earth. The extreme heat makes the existence of water and organic compounds impossible at the surface. Thus, the surface is not a likely habitat for life. Perhaps the clouds are a haven for organisms; there is believed to be a cloud layer with temperatures of only 15° to 32° Celsius that contains some water and oxygen. If life exists there, it would require a completely airborne ecology, to which there is no counterpart.

LIFE ON MARS?

Mars is the fourth planet from the sun. Its diameter is only half of the earth. But it has many earthlike features, and these led astronomers to believe that Mars could support life. Some astronomers even thought intelligent beings lived on Mars.

These ideas had to be changed after unmanned probes flew past Mars and sent information back to earth. They revealed Mars as a cold, dry world, more like the moon than like the earth; vast areas are pitted with craters. The density of Mars's atmosphere is incredibly low, like the earth's atmosphere at an altitude of 50 kilometers. The atmosphere consists mostly of carbon dioxide, with a trace of water. Because the air is so "thin," ultraviolet radiation from space, which is harmful to most living things, easily reaches the Martian surface.

Water may once have covered large parts of the surface of Mars. The planet's polar caps are now known to be composed of frozen water, and soil samples taken by the Viking probes yielded water upon heating. The average surface temperature of Mars is about 22° Celsius lower than that of the earth. At the Martian equator, the temperature may reach 24° Celsius during the day; at night it may drop to -73° Celsius.

These conditions make Mars an unlikely habitat for life as we know it. But exobiologists think that some forms of life may exist there. Several factors support this possibility. For example, ultraviolet radiation is easily blocked by physical barriers. Oxygen, nitrogen, and water may be present in amounts just sufficient to support hardy organisms such as bacteria, molds, and lichens. The presence of a layer of permanently frozen water, or *permafrost*, in the subsurface soil seems possible. This ice may melt locally from time to time, thus providing water for living things. Nitrogen may exist in small quantities in the atmosphere; in the soil there may be chemical salts containing nitrogen.

Exobiologists also base expectations of Martian life on certain laboratory experiments. They have subjected bacteria, algae, molds, lichens, some higher plants, and insects to artificial Martian environments. A surprising number of these organisms survived; certain bacteria and molds even did well under the harsh conditions as long as water was present.

Other experimenters have produced carbon compounds under simulated Martian conditions. Mixtures of carbon monoxide, carbon dioxide, and water vapor in sterile soil or powdered glass were exposed to ultraviolet radiation. Several more-complex carbon-containing compounds resulted, compounds known to be involved in the formation of still-more-complex biological compounds. This experiment demonstrates that methane-ammonia-hydrogen atmospheres are not always necessary for the chemical evolution of life.

But exobiologists consider such experiments with caution. It is impossible to duplicate Martian conditions exactly. Also, the test organisms in the exobiological experiments were given enough water and oxygen on which to survive.

The two Viking probes did answer several important questions. The Martian soil samples contained no organic matter, not even trace elements of carbon from meteorites that must have impacted on the planet. Although no life-forms were discovered, sensitive instruments detected an un-

usual chemistry that simulated several life-like reactions, including the breakdown of certain nutrient chemicals and the synthesis of organic compounds from gases.

OUTER PLANETS AND LIFE

For many years Jupiter, Saturn, Uranus, and Neptune have not been considered to be suitable homes for life. If we examine the facts about these planets, this attitude is not surprising.

Because of their great distances from the sun, the outer planets are extremely cold. Astronomers are not certain that all the outer planets have solid bodies. In spite of their immense sizes and masses, Jupiter, Saturn, Uranus, and Neptune are less dense than the earth. If present, the solid core of an outer planet may be small compared with the entire volume. It would probably consist of liquid and frozen gases, metals, and perhaps rock.

You can see why scientists find it difficult to visualize the existence of life on the outer planets. But today, for several reasons, exobiologists are changing their views. First, they now know that ammonia-methane-hydrogen atmospheres can be breeding places for biochemicals and thus of great importance to chemical evolution. Moreover, the outer planets are not always as cold as astronomers once thought. Instrumental studies of these worlds show that some of their cloud layers may be comparatively mild in temperature.

Jupiter has particularly been singled out for attention. It may have an internal source of heat energy that makes it warmer than it would be if heated by the sun alone. Its many-colored cloud coat may contain a large variety of organic compounds. Laboratory experiments have shown that organic and biologic compounds may well be in production in the atmosphere of Jupiter today. During Pioneer 10's historic flyby, earlier reports that Jupiter is a liquid planet were confirmed. Violent lightning flashes also were detected in Jupiter's clouds.

Pluto, the outermost planet of our solar system, is of little exobiological interest. Very little is known about it. It seems to be a small, solid body, intensely cold because

of its vast distance from the sun. Any atmosphere it could have would be permanently frozen.

Other solar system bodies may be of exobiological importance. We have already discussed meteorites. Asteroids, which probably are an important source of meteorites, may contain organic chemicals. Comets are also of great interest, and spectroscopic studies have already revealed carbon-bearing molecules in comets.

EXOBIOLICAL EXPLORATION

Exobiological knowledge of the planets and possibly other bodies in our solar system will increase as more spacecraft carry scientific instruments and men into space. Unmanned space probes have already flown past, orbited, and landed on Mars and Venus. These craft, however, were not designed, and otherwise failed, to detect the presence of living things on these two planets. Nor has any trace, past or present, of life on the moon been unmistakably found.

Future spacecraft that will land on Venus and especially Mars will carry instruments that should be able to identify biological chemicals and possible life-forms. Other craft have been sent toward Mercury and also the outer planets. The Jupiter probe, Voyager 2, is expected to pass close enough to Uranus in January 1986 and to Neptune in 1989 to gather and transmit back to earth a great deal of important information about these outer planets.

CIVILIZATIONS IN OUTER SPACE

It would indeed be a triumph to find any sort of life on another planet. But it would be even more exciting to discover intelligent life. How would one go about contacting intelligent beings on other worlds? The idea of communicating with creatures on other planets arose when man first realized that the planets are distinct worlds in space. Early telescopic astronomers possessed very little knowledge of conditions on the different planets of the solar system. They therefore imagined that these planets were inhabited by intelligent beings.

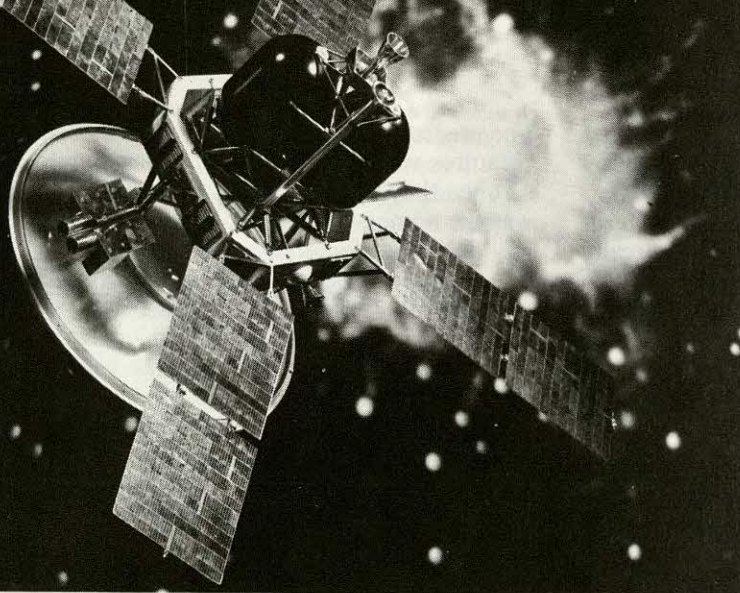
The astronomers reasoned that these creatures would wish to communicate with us once they found out that the earth, too, was the home of intelligent life.

Schemes of interplanetary communication were proposed from time to time. All of them involved the use of huge letters, symbols, numbers, or geometric figures that supposedly could be seen from the worlds nearest the earth: Mars, Venus, and the moon. For example, one man planned to have long ditches dug into the earth in the form of different signs. The ditches were then to be filled with flammable oil and set afire. Presumably, intelligent observers on neighboring planets would notice the fiery message and thus learn that the earth was civilized.

Such ideas were too impractical, and few persons took them seriously. Furthermore, the effort would have been wasted, for we now know that intelligent life probably does not exist in our solar system except on the earth.

With the advent of radio at the end of the 19th century, the idea of sending electronic messages or signals to other planets gained ground. Into the 1920's, signals were beamed into space. Occasionally, a radio operator claimed that he had received signals that could not have been generated by earthly stations. But most of these "messages" were probably caused by faulty equipment or by the electrical activities of the atmosphere. Others may have been radio waves from outer space. The existence of cosmic radio waves was proved in 1938. But to the best of our knowledge, these waves are of natural origin. (As we mentioned earlier, radio astronomers regularly tune in such cosmic waves to study the nature of the heavenly bodies that emit them.)

In the late 1960's, the world was astonished by news that electromagnetic pulses of extraordinary regularity were being received from certain locations in outer space. Newspapers and magazines published theories about cosmic supercivilizations that were broadcasting messages, or about alien spaceships that were being guided by space beacons. But astronomers



A model of the Viking spacecraft in simulated flight. Two of these probes have orbited and landed on Mars, taking many photographs of the planet's surface, performing atmospheric tests, and scooping up and analyzing soil samples for possible indications of life.

NASA

soon learned that the signals are natural; they are being emitted by unusual stars called pulsars.

Such pulses are just one instance of the many kinds of electromagnetic radiation in space. The radiation ranges from ultrashort gamma rays to very long radio waves. The task of sorting through all these wavelengths to find an intelligible message would be very difficult. Authorities suggest tuning in the wavelength emitted by hydrogen molecules (21 centimeters), which is the most common of the cosmic radio waves. This universal wavelength could be utilized by a civilization to send messages through space. But other wavelengths could serve just as easily.

Radio waves and other forms of electromagnetic radiation, such as light, move through a vacuum at approximately 300,000 kilometers per second. At this speed, the radiation will travel nearly 9.6 trillion kilometers in one year. This distance is called a light-year. To us on earth this seems an impossible distance. Nothing can be so far away! But the star nearest our solar system, Alpha Centauri, is 4.4 light-years from us. Thus, if we sent a radio message to a planet circling Alpha Centauri, it would take about 4.4 years to reach its destination. An answer would take at least that long to reach us. Thus, two-way communication between earth and an Alpha Centauri planet would involve nearly nine years for a single exchange of messages!

Many stars and their planets are millions or billions of light-years from the earth. Therefore, two-way communication with any intelligent beings on those worlds is impossible. By the time a message reaches us from one of these distant civilizations, the civilization may long since have passed out of existence. But simply to receive an intelligible message or signal from outer space would be exciting. It would prove that intelligent beings do exist, or existed, somewhere else in the universe.

ARE WE ALONE?

The search for extraterrestrial intelligence (SETI) has been compared to finding a needle in the "cosmic haystack." In spite of the enormous difficulties involved, scientists have attempted to listen for intelligent signals from outer space. In 1960-61 Dr. Frank Drake, at the National Radio Astronomy Observatory at Green Bank, West Virginia, used a radio telescope for this purpose. He directed the telescope's 26-meter antenna at two nearby stars that emit radio waves. This experiment was named Project Ozma. Dr. Drake failed to discover any intelligent patterns in the radio noise he received from these stars.

Similar experiments in the Soviet Union had no conclusive results. For a time, one investigator detected signals coming with some regularity from a part of the sky. But this event was not confirmed by other radio astronomers.

Scientists, however, did not give up hope. Since the end of Project Ozma, more than 20 radio-telescope searches have been conducted, mostly in the United States and the Soviet Union. A study named Project Cyclops was undertaken by the Ames Research Center of NASA and by Stanford University. Its purpose was to determine the possibilities of receiving messages from outer space. Project Cyclops suggested establishing a vast array of radio telescopes: up to 10,000 dish-shaped radio-telescope antennas, each 30 meters in diameter, would be spread over an area 16 to 32 kilometers wide.

Such an array would be able to detect radio and microwave leakages from planetary civilizations as much as 100 light-years from earth. Leakage results when some of the waves used for radio, television, and radar escape from a planet into outer space. This leakage may travel many light-years. Similar waves escape from the earth daily; it is possible that some distant civilization is even now tuning in the earth's electronic leakage.

Although NASA supported the SETI concept, federal funding became a serious problem in the United States. Senate action in 1981 withdrew financial support for NASA's involvement in the SETI effort. In December 1981, however, a SETI con-

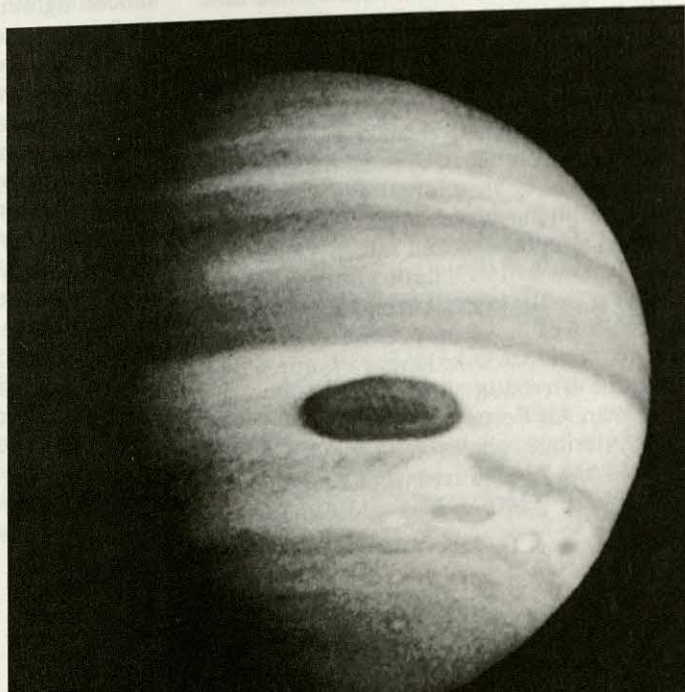
ference in Tallinn, U.S.S.R., enabled world scientists to establish plans for an international approach to the search for life in space. A further step in encouraging the search was taken in August 1982, when the members of the International Astronomical Union voted to create a commission on the search for extraterrestrial life.

Yielding to arguments that there might be useful by-products from the work, the U.S. government renewed some of its financial support for NASA's SETI involvement. That organization then made plans to use several of its large radio telescopes in the search, including its 91-meter Goldstone dish in the Mojave Desert and 300-meter Arecibo antenna in Puerto Rico.

In addition, new equipment enabled scientists to broaden their search efforts. At Harvard University in 1983, a four-year search for signals from space began, using a 25-meter-wide radio antenna. This was linked to a compact multichannel receiver that can monitor 131,072 channels simultaneously, and to a computer that can analyze a signal for the patterns that might mean a message.

Whether life, intelligent or not, exists on other planets remains to be learned. Many people dislike the thought of living things elsewhere in the universe. Other people welcome these possibilities.

Photo of Jupiter taken during Pioneer flyby of the planet. Exobiologists think that Jupiter's methane-ammonia-hydrogen atmosphere could possibly be producing biological chemicals.



NASA

UNIDENTIFIED FLYING OBJECTS

On June 24, 1947, a Boise, Idaho, businessman, Kenneth Arnold, was flying a private plane near Mount Rainier, Washington. Suddenly he was startled to see a group of strange-looking craft going through a series of amazing maneuvers. "They flew very close to the mountain tops," he said later, "flying . . . as if they were linked together . . . I watched them for about three minutes—a chain of saucer-like things at least five miles long, swerving in and out of the high mountain peaks. They were flat like a piepan and so shiny they reflected the sun like a mirror. I never saw anything so fast."

When Arnold reached his destination—Yakima, Washington—and reported what he had seen, he created a sensation. The "flying saucers" sighted by Arnold became an exciting topic of conversation. Within a few days disklike flying craft were reported by observers in other parts of the country, and the flying-saucer scare was fairly launched.

The scurrying disks that were now so frequently reported were merely the latest in a series of mysterious objects in the heavens that have startled men since time immemorial. Most of the earlier appearances had been definitely traced later to meteors, comets, atmospheric phenomena, and the like; a few have never been satisfactorily explained. Some aroused widespread interest for weeks or months at a time; then they were forgotten.

The flying-saucer epidemic started by Arnold's report reached rather respectable proportions in 1947; but toward the end of the year public interest seemed to be on the wane. It was powerfully revived by a tragedy that took place on January 7, 1948. Early in the afternoon of that day, observers at Godman Air Force Base, in Kentucky, saw a mysterious object flying overhead; it looked like "an ice cream cone topped with red." Four pilots in National Guard F-51

planes were asked to investigate the strange aircraft. Captain Thomas F. Mantell, the flight leader, radioed to the control tower that he was "closing in to take a good look." After a time he reported that the thing looked metallic and of tremendous size. "It's going up now and forward as fast as I am. . . . That's 360 miles per hour. I'm going up to 20,000 feet and if I am no closer, I'll abandon chase."

Following this last report, received at 3:15 P.M., there was no further radio contact with Mantell. Later that day his body was found in the wreckage of his plane near Fort Knox. The official U.S. Air Force explanation was that Mantell had blacked out at 6,000 meters (20,000 feet) from lack of oxygen and had died of suffocation before the crash. The object that Mantell had pursued was at first identified with the planet Venus; but further probing showed that the planet did not appear on that day in the quarter of the sky where the mysterious object had been sighted.

MANY SIGHTINGS REPORTED

There was now a new wave of flying-saucer sightings. Some of the mysterious craft were seen by planes, others by observers on the ground; a number of them were picked up on radar screens. Those making the reports included various seasoned observers: trackers of guided missiles, radar operators, commercial air pilots, U.S. Air Force pilots, airport traffic controllers, weather observers. The saucers were sighted in many different parts of the United States, but particularly in the desert areas of the Southwest. There were also reports of sightings from Canada, Mexico, South America, Europe, the Far East, Australia, Africa, Hawaii, Greenland, and the Antarctic.

The U.S. Air Force, determined to get to the bottom of these mysterious appearances, launched a series of investigations.



A Peruvian architect photographed this object in a valley near Lima, Peru. Was it a ship from outer space? A hoax? Or was it caused by normal phenomena?

UPI

The investigators gave the official name of "unidentified flying objects" (UFOs) to the mysterious craft.

The UFOs reported by observers were of many different kinds. Flying saucers, or disks, predominated. There were large disks, up to 30 meters or so in diameter, medium-sized disks, and tiny disks with a diameter of only a few centimeters. These craft performed amazing maneuvers. They sometimes hovered motionless in the air, then shot skyward, made abrupt turns, and reversed their course with unbelievable suddenness. They attained the most fantastic speeds—up to thousands of kilometers an hour, in some cases.

There were also rocket-shaped ships, ranging from 30 to 300 meters in length and also capable of tremendous speeds. In some cases these rockets seemed to serve as mother ships for disk-shaped craft. In addition to disks and rockets, there were bright green fireballs moving silently through the heavens as swiftly as meteors.

ALIEN BEINGS SEEN

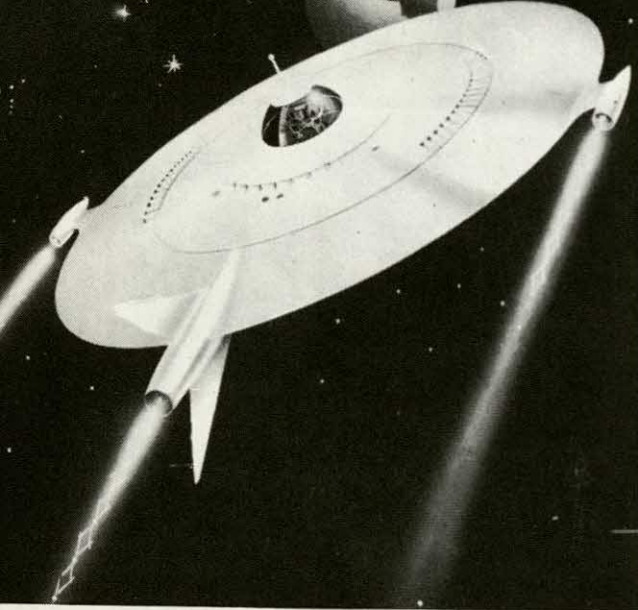
Certain observers claimed to have seen members of the flying-saucers' crews. In a 1950 book called *Behind the Flying Saucers*, Frank Scully told of certain men from Venus whose dead bodies had been found after their saucer craft had crashed to earth in New Mexico and Colorado.

According to Mr. Scully's informant, a "Mr. Newton," the Venus men were "tiny creatures, from 97 to 112 centimeters long. They wore 1890 dress, of cloth that was not wool or cotton but that couldn't be torn." The Air Force, according to Mr. Newton, had spirited the bodies away.

Not all UFO crew members were midgets, apparently. On September 12, 1952, a woman, three children and a young National Guardsman espied a flying saucer near Sutton, West Virginia. Somewhat later they came upon a repulsive giant, apparently a member of the crew. He was 2.7 meters tall and had a red face and protruding eyes about 30 centimeters apart. When this monster started toward the startled observers with a hissing sound, they fled.

There were saucer photographs galore. Some of these were obviously hoaxes; in other cases freak images had been produced in negatives because of faulty equipment or technique. A few could not be explained so easily, however.

Flying-saucer appearances tapered off after 1953. Perhaps it would be more exact to say that fewer reports of unidentified flying objects appeared in newspapers and magazines, perhaps because editors felt that such reports were no longer particularly newsworthy. In the period 1965–67, there was a marked increase in reported sightings and then a tapering off again.



United Press

An artist's conception of a flying saucer, based on descriptions given in early reported sightings.

LOOKING FOR EXPLANATIONS

How are we to account for these mysterious, unidentified objects in the heavens? Are they really flying craft, incomparably more advanced than any previously known to man? After all, it has been argued, man has traveled to the moon. Is it not possible that inhabitants of other star systems, further advanced than man, have perfected ships that can carry them to our solar system?

Soon after Kenneth Arnold sighted those "flying saucers" in 1947, the U.S. Air Force began keeping track of all reported UFO sightings in what came to be known as Project Blue Book. In 1966 they engaged a group of 36 scientists, under the direction of physicist Edward U. Condon of the University of Colorado, to conduct an independent investigation of these sightings. The study culminated in the publication of a voluminous report, which concluded that further UFO research would be of no value. In 1969 the Air Force ended Project Blue Book, although approximately 1,000 reports in their files remained unexplained.

Reports of UFO sightings dropped sharply, but the controversy over their cause didn't end. Some scientists continued to question the thoroughness of the Air

Force's investigation. Dr. James McDonald, for example, noted that the authors of the Condon Report "conceded that about one third of the 90 cases they investigated could not be explained. It is far from clear how this justified the conclusion that further study is not needed."

Then in 1973 there was a sudden upsurge in UFO sightings. One of the more interesting accounts was related by two men from Mississippi, who said that they had been taken aboard a blue spaceship by creatures who had silvery, wrinkled skin and crab-claw hands. Astronomer J. Allen Hynek, who had been an Air Force consultant on Project Blue Book, examined the men under hypnosis and felt that they were "telling the truth beyond a reasonable doubt."

Other sightings soon proved to have conventional explanations. For example, a number of UFOs were actually silvery weather balloons, which reflected sunlight at their 26 kilometers altitude. Others turned out to be meteorological phenomena such as low-hanging clouds or fog. Nonetheless, a survey of adult Americans showed that 51 per cent believed that UFOs are real objects, not just people's imaginings. And 11 per cent said that they had seen UFOs.

Late in 1973, Dr. Hynek, who is now Chairman of the Department of Astronomy at Northwestern University, announced the formation of the Center for UFO Studies. "For a quarter century the UFO phenomenon has been the subject of gross misconceptions, misinformation, ridicule, buffoonery, and unscientific approach," he said. "The fact that reports persist—from many countries—presents a mystery that demands explanation."

SOME CONVENTIONAL EXPLANATIONS

As has been indicated, many flying saucers are not really flying craft at all. Many scientists, such as astronomer Donald H. Menzel of Harvard, maintain that many of the mysterious appearances have been due to meteorological phenomena. In some cases, saucerlike appearances may be due to the reflection of light from ice crystal formations in the atmosphere. If such crys-

tals are falling or hovering in the air, reflected sunlight or moonlight may produce startling effects. For example, such reflections may cause a pair of concentric halos to appear around the sun. Sometimes part of the halos may be almost as bright as the sun itself and will form glowing mock suns, which are also known as sundogs. At night, halos sometimes develop around the moon, forming mock moons, or moondogs. These might well resemble strange aircraft.

In other instances, flying-saucer appearances may have been due to the fact that air can act as a distorting lens, producing the effects known as mirages. Light rays are refracted, or bent, in various ways as they pass through air layers of different density. The effect is particularly noticeable when layers of sharply contrasting density are in close contact with one another.

If a hot layer of air lies close to the earth, the image of the sky may be projected against the earth, sometimes giving the illusion of ponds or lakes in the distance. If a cool layer is near the earth (as in the desert at night), the image of the earth may be projected against the sky. Distant lights, such as those of automobiles or a city, will seem to float in the air. If the air is turbulent, these lights will apparently dart hither and thither. Conditions that cause optical mirages can also produce radar mirages.

In some cases, people may have sighted flying craft, but not the exotic objects that they imagined. They may have seen ordinary airplanes flying at such great heights that only the reflection of the sun from the fuselage was visible.

Other people have mistaken kites or weather balloons for UFOs. In 1951, physicist Urner Liddel asserted that most UFOs previously reported were really skyhook balloons—large plastic unmanned craft, which are used to carry meteorological instruments aloft. These balloons reach great heights and often fly great distances; some have crossed the Atlantic. When viewed against the background of the sky, they often look quite disklike. Some authorities believe that it was while chasing a skyhook balloon that Captain Mantell crashed. It was a secret device at the time and infor-

mation about it was not available to the public.

When this information was declassified, a few years later, it solved a startling mystery. On March 17, 1950, the inhabitants of Farmington, New Mexico, were almost scared out of their wits as thousands of "flying saucers" soared over the town for an hour or so. Only later was it revealed that a skyhook balloon had burst over Farmington at an altitude of 18 kilometers scattering thousands of pieces of plastic in the air. These were the "flying saucers" that had invaded the area.

PRODUCTS OF THE MIND?

Behavioral scientists also study UFO reports. Sociologist Robert Hall points out that "the sky, especially the night sky, is full of ambiguous stimuli, and people generally have a powerful need to reduce ambiguity . . . by explanations in terms of something familiar." Thus some people, influenced by the general "system of belief" that has developed around the UFO phenomenon, transform naturally explainable stimuli into a ship from outer space.

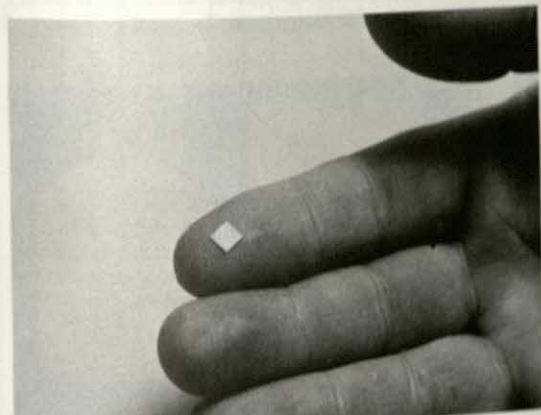
Dr. Lester Grinspoon suggests that UFO sightings may be linked to the stresses of modern life—to "the increasingly anxious times in which we live." Under such stresses, he says, it is common for people to have illusions or delusions, substituting fantasy "to supply what reality has denied."

Dr. Carl Sagan of Cornell University has proposed that "certain psychological needs [are] met by belief in superior beings from other worlds." And today, he says, extraterrestrial visitors are a fashionable idea. Dr. Sagan, however, cautions against dismissing the extraterrestrial hypothesis. There is not enough evidence, he believes, to exclude the possibility that some UFOs are spaceships from advanced civilizations that live elsewhere in the universe.

In 1794, France's Académie des Sciences refused to accept the possibility that meteorites originated outside of the earth's atmosphere. "Meteorites didn't fit established belief, so scientists denied their existence," said Dr. Hall. Is the same true for flying saucers from outer space?



COMPUTERS & MATHEMATICS



Left: Sperry Univac division of Sperry Rand Corporation
Above: General Automation Inc.

Computers have become an integral part of modern life. Left: a full-size computer with central processing unit (background), storage unit—magnetic disks (upper left), and output, or printer, unit (foreground). Above: the data processing chip of a modern microcomputer.

| | |
|---------|-----------------------------|
| 332-337 | Introduction to Mathematics |
| 338-345 | Numerals |
| 346-351 | Arithmetic |
| 352-365 | Algebra |
| 366-379 | Plane Geometry |
| 380-385 | Solid Geometry |
| 386-390 | Trigonometry |
| 391-400 | Analytic Geometry |
| 401-403 | Non-Euclidean Geometry |
| 404-413 | Statistics |
| 414-423 | Probability |
| 424-430 | Game Theory |
| 431-437 | Calculus |
| 438-445 | Set Theory |
| 446-452 | Binary Numerals |
| 453-461 | Data Processing |
| 462-474 | Computers |
| 475-483 | Computer Programming |

INTRODUCTION TO MATHEMATICS

The word "mathematics" comes from the Greek *mathemata*, meaning "things that are learned." It may seem odd to apply this phrase to a single field of knowledge, but we should point out that for the ancient Greeks, mathematics included not only the study of numbers and space but also astronomy and music. Nowadays, of course, we do not think of astronomy and music as mathematical subjects; yet the scope of mathematics today is broader than ever.

Modern mathematics is a vast field of knowledge with many subdivisions. There is, first of all, the mathematics of numbers, or quantity. The branch of *arithmetic* deals with particular numbers, such as 3, or $10\frac{1}{2}$ or 12.5. When we add, subtract, multiply, or divide such numbers or get their square roots or squares, we are engaging in arithmetical operations. Sometimes we wish to consider, not particular numbers, but relationships that will apply to whole groups of numbers. We study such relationships in

algebra, another branch of the science of quantity. In algebra, a symbol, such as the letter *a* or *b*, stands for an entire class of numbers. For example, in the formula

$$(a + 2)^2 = a^2 + 4a + 4$$

the letter *a* represents any number. The relationship expressed in the formula would remain the same whether *a* stood for 1, or 5, or 10 or any number.

Mathematics also studies shapes occurring in space, which may be thought of as a world of points, surfaces and solids. We study the properties of different shapes and the relations between them, and we learn how to measure them. This space science is called *geometry*. *Plane geometry* is concerned with points, lines, and figures occurring in a single plane—a surface with only two dimensions (Figure 1). The study of the three-dimensional world is called *solid geometry* (Figure 2). Trigonometry ("triangle measurement") is an offshoot of

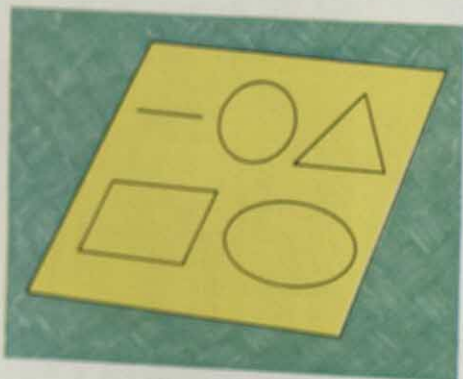
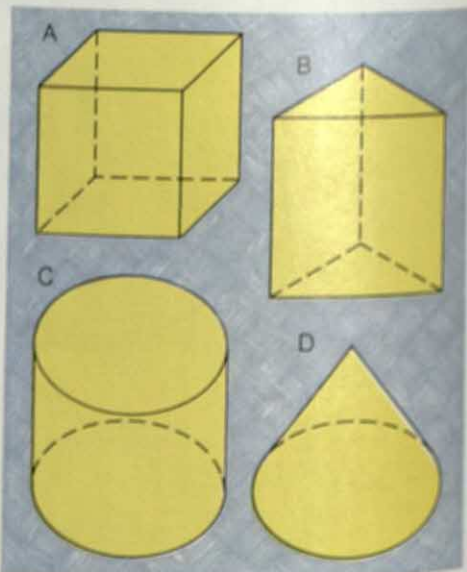


Figure 1. The straight line, circle, triangle, rectangle and ellipse shown above occur in a single plane. Their study is a part of plane geometry.

Figure 2. The cube (A), prism (B), cylinder (C), and cone (D) are not bounded by a single plane. These figures are studied in solid geometry.



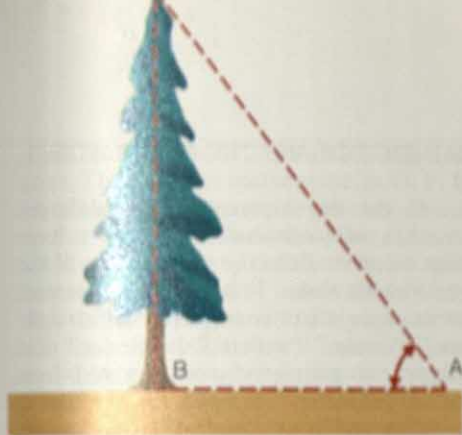


Figure 3. If we know the distance AB and the angle at A , we can calculate how high the tree is by using trigonometry, or triangle measurement, an offshoot of geometry.

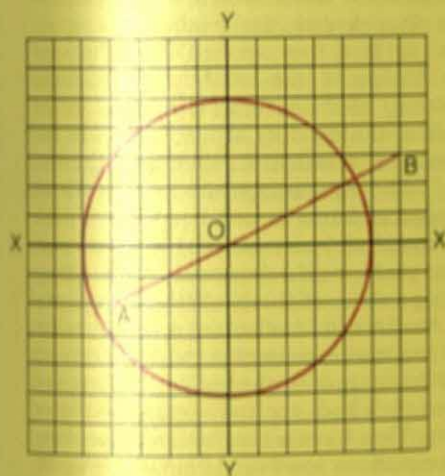


Figure 4. The line XX' is at right angles to YY' . The formula $x = 2y$ indicates the position of AB with respect to XX' and YY' . The formula $x^2 + y^2 = 25$ indicates the position of the circle. We are dealing here with analytic geometry.

geometry. It is based on the fact that when certain parts of triangles are known, one can determine the remaining parts and solve many different problems (Figure 3).

Analytic geometry combines algebra and geometry—generalized numbers and space relationships. It locates geometrical figures in space. It explains circles and ellipses and other figures in terms of algebraic formulas. In Figure 4, for example, the lines XX' and YY' are at right angles to one

another and meet at the point O . To indicate the position of the straight line AB with respect to XX' and YY' , we use the formula $x = 2y$. To indicate the position of the circle in the diagram, we use the formula $x^2 + y^2 = 25$.

The branch of mathematics called *calculus* is based on the study of functions. If the value of a given quantity depends on the value we assign to a second quantity, we say that the first quantity is a function of the second. For example, we know that the circumference of a circle is always 3.14159 times the diameter. The value 3.14159 is indicated by the Greek letter π , or pi. We frequently use the formula $C = \pi D$ (that is, $\pi \times D$), where C is the circumference and D the diameter. In this case, C , the circumference, is a function of D , the diameter, since its value will depend on the value we assign to D . In *integral calculus*, we are interested in the limit of the different values of a variable function. In *differential calculus*, we determine the rate of change of a variable function.

Statistics, another branch of mathematics, involves the accumulation and tabulation of data expressed in quantities, and the setting up of general laws based on such data. The theory of probability is an important part of statistics. It enables one to calculate what will happen when the chances of an event occurring are even and also when the chances are not even. The theory has been applied to gambling games, such as dice and poker. It can be used, too, to predict the relative number of boys and girls that will be born in a particular place at a particular time.

These are only a few of the many subdivisions of mathematics. Besides being a most extensive field of knowledge in its own right, it represents a logical approach that can be applied to many different fields. It carefully defines the ideas that are to be discussed and clearly states the assumptions that can be made. Then, on the basis of both the definitions and the assumptions, it forges a chain of proofs, each link in the chain being as strong as any other. Mathematicians have displayed wonderful powers

of imagination in determining what can be proved and in constructing ingenious methods of proof.

It might seem rather far-fetched to think of mathematics as a search for beauty. Yet, to many workers in the field, mathematical patterns that are fitted together to form a harmonious whole can produce as pleasing effects as the color combinations of a painter or the word patterns of a poet. As Bertrand Russell, who was an important 20th century mathematician, has put it in his *Principles of Mathematics*: "Mathematics, rightly viewed, possesses a beauty cold and austere, like that of a sculpture, without any appeal to any part of our weaker nature, without the gorgeous trappings of painting or music, yet sublimely pure, and capable of a stern perfection such as only the greatest art can show."

Mathematics is also an endless source of entertainment. For many generations, mathematicians and others have prepared what are commonly known as mathematical recreations, ranging from simple problems and constructions to brain twisters that can be solved only by experts—and sometimes not even by experts. These recreations are a delightful challenge to one's wits. Sometimes, they bring us into a world of fantasy in which one "proves" that $2 = 1$, or constructs a magic ring whose outside is its inside (Figure 5).

MATHEMATICS AND THE OUTER WORLD

In the development of the different branches of mathematics, pioneers have often owed much to the observation of the world about them. It has been suggested, for example, that the concepts of "straight line," "circle," "sphere," "cylinder," and "angle" in geometry were derived from natural objects: "straight line," from a towering tree trunk; "circle," from the disk of the sun or moon; "sphere," from a round object like a berry; "cylinder," from a fallen tree trunk; "angle," from the angles formed by the arms or legs in varying positions.

Pioneer mathematicians examined these shapes and the relations between them. At first, they applied the results of such studies to the solution of practical problems, such as the construction of canals or the dividing of land into lots for purposes of taxation. Later, they began to study the relationships between various geometrical forms in order to satisfy their intellectual curiosity and not to solve particular problems. In the course of time they built up a series of purely abstract concepts.

As certain mathematicians develop such concepts in geometry and calculus and other branches of mathematics, they are apt to turn their backs entirely on the world of reality. They have no hesitation in working with equations involving four dimensions,

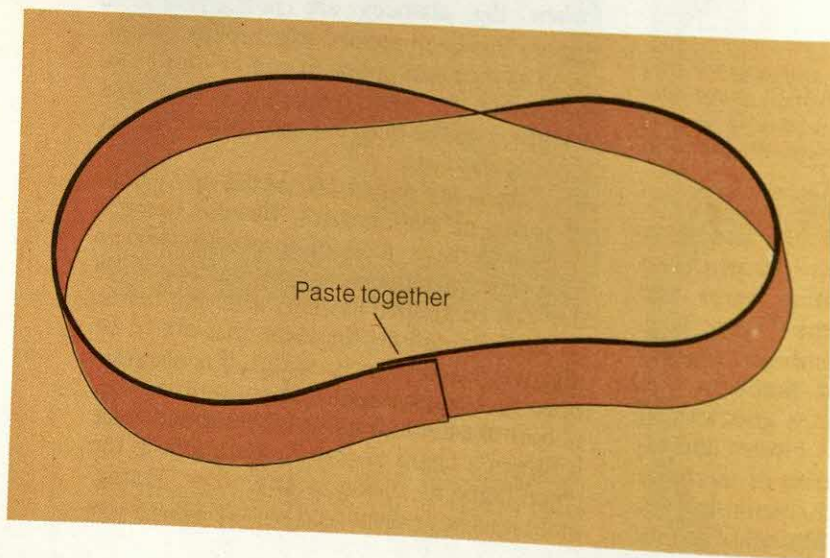


Figure 5. This magic ring is made by pasting together the two ends of a narrow strip of paper in the form of a ring, after giving one end a twist of 180° . If you then color one side of the ring, you will find that you have colored the whole ring, inside and out. The ring is called a Moebius Strip.

or ten dimensions, or any number of dimensions. The German mathematician G. F. B. Riemann built up an entire geometry based on the postulate, or assumption, that no two lines are ever parallel. Another geometry—that created independently by N. I. Lobachevsky and J. Bolyai—started with the assumption that at least two lines can be drawn through a given point parallel to a given line. The German Georg Cantor proposed a “theory of sets” that was highly ingenious and—at least, so it appeared at the time—utterly useless as far as any practical application was concerned.

It has been pointed out that the concern with pure abstraction is not without its dangers. As the distinguished mathematician Morris Kline has put it in his *Mathematics and the Physical World*: “Mathematicians may like to rise into the clouds of abstract thought, but they should, and indeed they must, return to earth for . . . food or else die of mental starvation.”

Yet even the most abstract speculations of mathematicians may find important applications, sometimes after many years have passed. Thus the Riemannian geometry was to prove invaluable to Albert Einstein when he developed his theory of relativity. Georg Cantor’s theory of sets has been applied to various fields, including higher algebra and statistics. As to multiple dimensions, they have been put to work, of all things, in the inspection of television tubes.

APPLICATIONS

We do not have to give such extreme examples as Riemannian geometry and the theory of sets to show how mathematics has served mankind. Actually, it is deeply rooted in almost every kind of human activity, from the world of everyday affairs to the advanced researches of authorities in many different fields of science.

All of us are mathematicians to some extent. We use arithmetic every day in our lives—when we consult a watch or clock to find out what time it is; when we calculate the cost of purchases and the change that is due us; when we keep score in tennis, or baseball, or football.

The accounting operations of business and industry are based on mathematics. Insurance is largely a matter of the compounding of interest and the application of the theory of probability. Certain manufacturers make use of calculus in order to be able to utilize raw materials most effectively. The pilot of a ship or plane uses geometry in plotting his course. Much of surveying is based on trigonometry. The civil engineer uses arithmetic, algebra, calculus, and other branches of mathematics in his work.

Mathematics also serves the branches of learning called the humanities, which include painting and music. It is the basis of perspective—the system by which the artist represents on a flat surface objects and persons as they actually appear in a three-dimensional world. We may think of perspective as made up of a series of mathematical theorems. One theorem, for example, states that parallel horizontal lines that recede in the same plane from the observer converge at a point called the vanishing point. Figure 6 shows how this theorem is applied to the drawing of a room.

In music, too, mathematics plays an important part. The system of scales and the theories of harmony and counterpoint are basically mathematical; so is the analysis of the tonal qualities of different instruments. Mathematics has helped greatly in the design of pianos, organs, violins, and flutes, and also of such reproducing devices as phonographs and radio receivers.

Mathematics is so important in science and serves in so many of its branches that it has been called the “Queen and Servant of the Sciences” by the noted Scottish-American mathematician, Eric Temple Bell. Here are some examples:

Measurements and other mathematical techniques are vital in the work of the physicist. The physicist uses the mathematical device called the graph to give a clear picture of the relationship between different values—between temperature, say, and the pressure of saturated water vapor in the atmosphere (Figure 7). The laws of physics are stated in the form of algebraic formulas. Thus to express the idea that the velocity of a body can be determined by dividing the

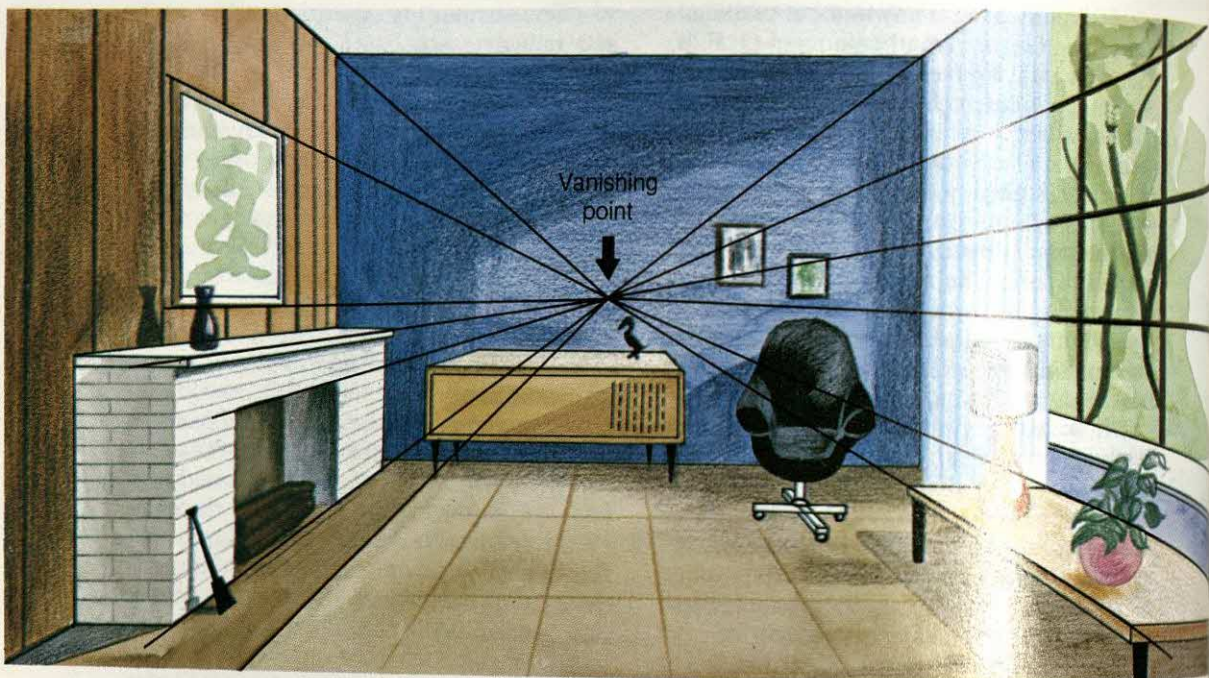


Figure 6. The realistic, three dimensional effect in the above drawing is obtained by having all the horizontal lines that recede from the observer meet at a vanishing point.

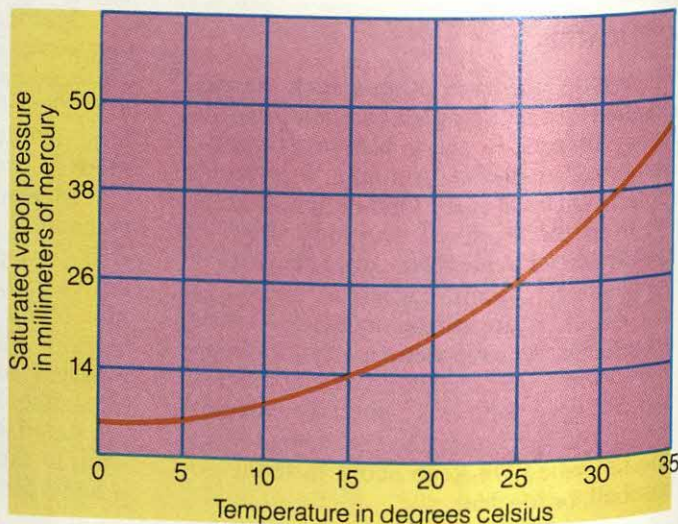
distance covered by the time required to cover this distance, the physicist uses the

formula $v = \frac{s}{t}$, where v is the velocity, s

the space or distance covered and t the time. The physicist employs geometry and trigonometry in his analysis of forces and in establishing the laws of optics, the science of light.

Like the physicist, the chemist continually uses arithmetical and algebraic operations and graphs in his work. The chemist too presents laws in the form of algebraic formulas. The reactions are set down in the

Figure 7. The graph at the right shows how saturated water-vapor pressure in the atmosphere varies with the temperature. The saturation point is reached when the atmosphere can hold no more water vapor at a given temperature.





Courtesy of Borg-Warner

The scope of mathematics today is broader than ever, and more sophisticated tools are being developed to assist students.

form of equations, which from certain viewpoints may be considered as mathematical equations. Chemists also use logarithms, a mathematical technique, in calculating the degree of acidity of a substance—the so-called pH value. Plane and solid geometry are used in studying the way in which atoms or ions (electrically charged atoms) are combined. Thus it can be shown that graphite atoms form a succession of hexagons (six-sided figures) in a series of planes set atop one another; and that the sodium and chloride ions that make up ordinary table salt (sodium chloride) are set at the corners of a series of cubes.

Mathematics has always been closely associated with astronomy. From the earliest days, astronomers measured angles and arcs and made a great many mathematical calculations as they followed the apparent motions of the sun, stars, moon and planets in the heavens. Today such branches of mathematics as arithmetic, algebra, plane geometry, solid geometry, trigonometry, and calculus are just as useful to the astronomer as the optical telescope, camera, radio telescope, and other devices that he uses in his work.

It would seem difficult to apply the formulas of mathematics to the infinitely

varied world of living things. Yet mathematics serves even in biology, the science of life. It plays an extremely important part, for example, in genetics, which is concerned with heredity. To calculate the percentage of individuals with like and unlike traits in succeeding generations, the geneticist makes use of the theory of probability. Mathematics has also been applied to the comparison of related forms of life. Using the method called dimensional analysis, researchers have found that the frequency of the beating of a bird's wings can be summed up in a formula. It is rather amazing to see how often this formula applies to totally dissimilar birds. Dimensional analysis has also been used to analyze the growth patterns of certain animals, as well as to determine the ratio between the lifetime of a given animal and the time required for the animal to draw a single breath.

Such then, in brief, is the scope of mathematics. In the articles that follow, we shall present some of the more important fields of mathematics, including arithmetic, algebra, plane and solid geometry, trigonometry, analytical geometry and calculus. We shall tell you what these fields are about and the uses to which they are put by present-day mathematicians.

NUMERALS

by Howard F. Fehr

In the Ashmolean Museum in Oxford, England, there is an Egyptian royal mace, on which there is a record of 120,000 prisoners and of booty consisting of 400,000 oxen and 1,422,000 goats. This record, dating back to before the year 3400 B.C.,

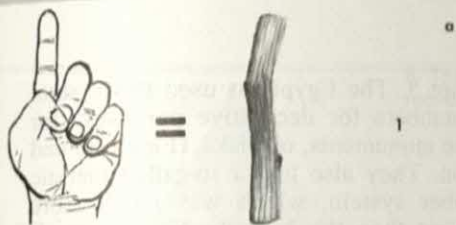
shows that in antiquity people had learned to write large numbers. Of course, the beginnings of the use of numbers must go back long before the Egyptians.

The primitive cave man did not have to know much about counting or any other kind of mathematics to keep alive. Home was a cave already at hand; food grew on trees or plants or could be hunted with primitive weapons. However, when people began to collect animals into herds, and particularly when one family entered upon social relations with others, it became necessary to decide how much belonged to one person and how much belonged to a neighbor. It probably sufficed, at the outset, to use such concepts as a little, or some, or much. Later, however, it became necessary to have a more definite means of determining how much. People learned to count and this was the beginning of mathematics.

At first, a person might count the number of animals in a herd by placing a pebble on the ground or tying a knot in a rope for each animal. Each pebble in the growing heap or each knot in the rope would stand for a single animal. Later, a man might use his ten fingers in his calculations. We may surmise that when the ten fingers had been counted, a little stone would be set aside to represent this first ten; the fingers would then be used to count another ten. Another stone would be added to the first one; the fingers would be used to count another ten, and so on. When the stones in the pile would equal the number of fingers, they would represent ten tens. The pile of ten stones would then be taken away, and a larger stone would be put in its place to indicate "ten tens" or a hundred (Figure 1). Thus three large stones, seven small stones, and eight sticks, standing for eight fingers, would represent three hundreds, seven tens, and eight units—in other words, 378 (Figure 2).

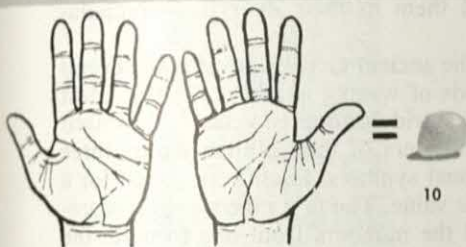
The ten fingers in this case would mark the halting place in a person's calculations:





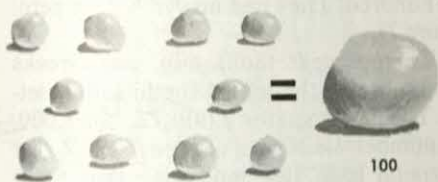
a

1



b

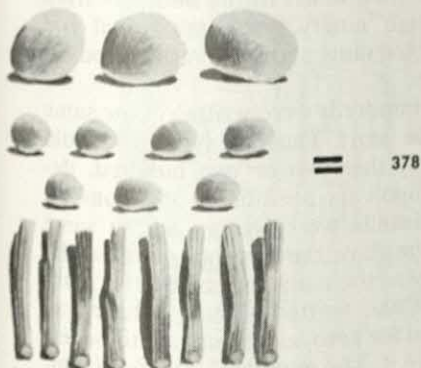
10



c

100

1. In the calculations of early times, the number one might be represented by a finger or by a stick (a). The number ten could be represented by the ten fingers of the two hands or by a small stone (b). Ten small stones would be equivalent to one large stone, which would represent one hundred (c). In this system, the ten fingers marked the stopping place in counting; in other words, the system was based on ten.



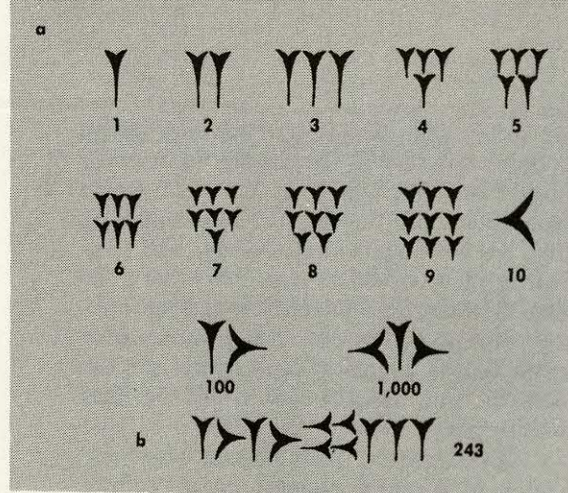
378

we would call it the base. Not all primitive peoples would use ten, or the number of fingers on both hands, as the base. Some would use only the two hands in counting (and not the fingers of the hands), and their halting place would be two. For others, the fingers of one hand would serve; their halting place would be five. Still others would combine the fingers of both hands and the toes of both feet; twenty would be their halting place.

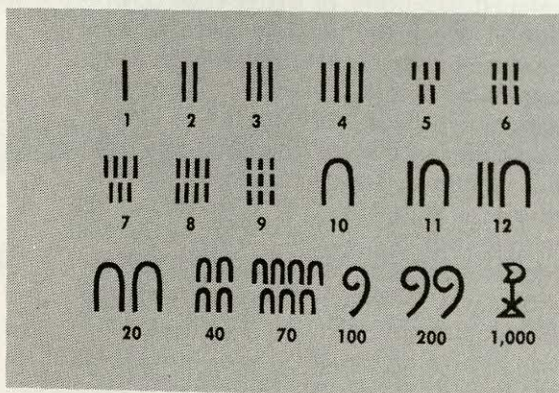
Since stones, pebbles, and sticks are awkward to handle, people created symbols to represent numbers as soon as they learned to write. The symbol that is used to write a number is called a *numeral*. About 5,000 years ago, the Babylonians developed cuneiform, or wedge-shaped, writing, in which the different symbols were produced with sharp-edged sticks on wet clay formed into flat bricks. There were special symbols for one, ten, and a hundred: ∇ stood for one; \angle , for ten; Υ , for a hundred. To write a number, the Babylonians repeated the symbols as required. In Figure 3a, we show how the first ten numbers were written and how 100 and 1,000 were written. Figure 3b gives the Babylonian equivalent for 243. Such a system of writing numbers is called additive, since we must add the symbols together to find the total number. To represent the higher numbers, the Babylonians used a multiplying procedure. For example, the symbol $\angle \Upsilon$ represented ten times one hundred; that is, one thousand. If the symbol of a smaller number preceded the symbol of a larger number, it indicated that the numbers were to be multiplied together. If the larger symbol came first, the numbers were to be added.

The ancient Egyptians had a purely additive system—the hieroglyphic system—which used the symbols given in Figure 4. The writing of numbers was a cumbersome business. The number 527, for example, would be written as indicated in

2. If we were to use the method of calculation illustrated in Figure 1, we would represent the number 378 by three large stones (representing 300), seven small stones (representing 70), and eight sticks (representing eight).



3. Babylonian cuneiform numerals: a: the symbols for numbers 1 through 10 and for 100 and 1,000; b: how the numbers 243 would be written.



4. The hieroglyphic numeral system used by the Egyptians, particularly for decorative purposes. They had another system—the hieratic system—for everyday computations.



5. The Egyptian hieroglyphic system was an additive system, and the writing of numbers became cumbersome. Above shows how the number 527 would be written.

Figure 5. The Egyptians used this system of numbers for decorative purposes—for stone monuments, obelisks, (Figure 6) and so on. They also had a so-called hieratic number system, which was much more efficient than the hieroglyphic system. It served them in their everyday computations.

The ancient Greeks developed several methods of writing numbers. In the latest and most widely adopted version, they used all the letters of their alphabet plus three additional symbols. Each letter stood for a definite value. The first nine symbols represented the numbers from one to nine; the next nine, the tens from ten to ninety; the last nine, the hundreds from one hundred to nine hundred. They had no symbol for zero. (Figure 7).

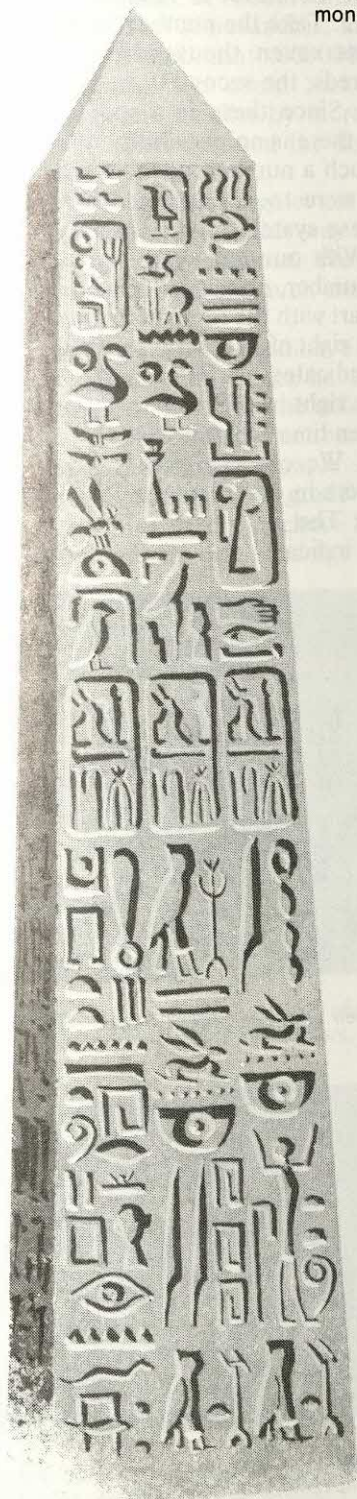
To represent thousands, the Greeks added a bar to the left of the first nine letters. Thus Γ stood for 3,000; Σ , for 7,000. The number 4,627 was written $\Delta \chi \kappa \zeta$, the horizontal base line indicating that these letters formed a numeral.

The Hebrews also used their alphabet in writing numbers (Figure 8).

A system of rodlike symbols was employed by the ancient Chinese to represent numbers. They had a place system: that is, a number symbol took on different values according to the place it occupied in the written number. In Figure 9, we give the numbers from one through nine and from ten through ninety. You will note that in the numbers from one through nine, each vertical rod stood for one unit and each horizontal rod for five units. In the numbers from ten through ninety, each horizontal rod stood for ten units and each vertical rod for fifty units.

The hundreds were written in the same way as the units. Thus the symbol Π would stand for either two or two hundred, depending upon its position in the number. The thousands were written in the same way as the tens, the ten thousands in the same way as the units and so on. The number 7,684 was written $\perp \tau \equiv \text{IIII}$. There was no symbol for zero, and a gap had to be left to indicate it. The number 7,004, for example, was \perp IIII. If the gap were not recog-

6. The Egyptians used the decorative hieroglyphic numerals in monuments such as the typical obelisk illustrated at left.



nized as such, the number might be read as 74 instead of 7,004.

The Romans probably derived their system of numbers from the Etruscans, earlier inhabitants of Italy. Letters were used for numerals. "I" stood for one, "V" for five, "X" for ten, "L" for fifty, "C" for a hundred, "D" for five hundred and "M" for a thousand. Two I's represented the number two; three I's, the number three. An I was put before V to make four. This was the application of the so-called subtractive principle: if the symbol of a smaller number preceded the symbol of a larger number, the smaller number was to be subtracted from the larger. VI stood for six; VII, for seven; VIII, for eight. The symbol for nine was IX (subtractive principle). Other numerals are shown in Figure 10.

So it was throughout the Roman numerical system. If a number was smaller than the number that followed it, it was subtracted from the second number; if it was as large as or larger than the following number, the second number was added to the first. Thus LX was sixty; XL was forty.

Roman numerals are still used in English and other Western languages for certain specific purposes. They sometimes indicate chapter numbers or volume numbers. They are also used on the dials of some clocks and also on commemorative monuments and tablets. To us, this number system seems extremely complicated. For example, we indicate the number eighteen hundred eighty eight by only four symbols: 1888. Thirteen symbols would be required in Roman numerals; it would be MDCCCLXXXVIII.

Our own system of numerals, the so-called Arabic numerals, should really be called the Hindu-Arabic numerals, since the system originated in India (not long before the Christian era) and was later adopted by the Arabs. The Arabs conquered a large part of Spain in the eighth century and in time introduced the Hindu-Arabic numerals in the conquered land. In Figure 11, we show how the numerals

looked in the tenth century. The system was gradually adopted by the other peoples of Europe. By the fifteenth century, the symbols of the system had acquired the form that is so familiar to us:

1, 2, 3, 4, 5, 6, 7, 8, 9, 0

The base of our system is ten, and so it is called a decimal system. (*Decem* means "ten" in Latin.) It is a place system, in which the position of a symbol indicates its particular value. Moving one space to the left increases the value tenfold, as is shown in the following diagram:

| | | | | |
|------------|----------|---------|--------|-----|
| 10,000's | 1,000's | 100's | 10's | 1's |
| (10×1,000) | (10×100) | (10×10) | (10×1) | |

In the number 4,962, the 4 stands for four thousands; the 9, for nine hundreds; the 6, for six tens; the 2, for two units.

4,962, therefore, is really $4,000 + 900 + 60 + 2$. Take the number 7,004. The 7 represents seven thousands; the first 0, no hundreds; the second 0, no tens; the 4, four units. Since there is a special symbol for zero, there is no possibility of error in reading such a number, as there would be if the zero were to be indicated, as in the early Chinese system, by leaving a gap.

With our ten symbols, we can write any number, no matter how large. Suppose we start with the number 1. If we put a zero to the right of it, the 1 is in the ten's column and indicates ten. Continuing to put zeros to the right in this way, we make the value of 1 ten times greater for every zero that we write. We can also write extremely small numbers in our system by using a decimal point. The first numeral after a decimal point indicates the number of tenths; the

| | | | | | | | | | | | | |
|----|----|----|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Α | Β | Γ | Δ | Ε | Ζ* | Η | Θ | Ι | Κ | Λ | Μ | Ν |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 20 | 30 | 40 | 50 |
| Ξ | Ο | Π | Ρ* | Σ | Τ | Υ | Φ | Χ | Ψ | Ω | Ϟ* | |
| 60 | 70 | 80 | 90 | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 |

7. The ancient Greeks wrote numbers using the letters of their alphabet plus three additional symbols, indicated above by asterisks.

| | | | | | | | | | | | | | |
|----|----|----|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|----|
| א | ב | ג | ד | ה | ו | ז | ח | ט | י | כ | ל | מ | נ |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 20 | 30 | 40 | 50 |
| ס | ע | פ | צ | ק | ר | ש | ת | ך | ם | ן | ף | ץ | |
| 60 | 70 | 80 | 90 | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | |

8. Like the Greeks, the ancient Hebrews used the letters of the alphabet in writing their numbers.

second, the number of hundredths; the third, the number of thousandths, and so on. .000001 is a millionth; .00000001, a hundred millionth. We could make the number smaller and smaller by increasing the number of zeros between the decimal point and the 1.

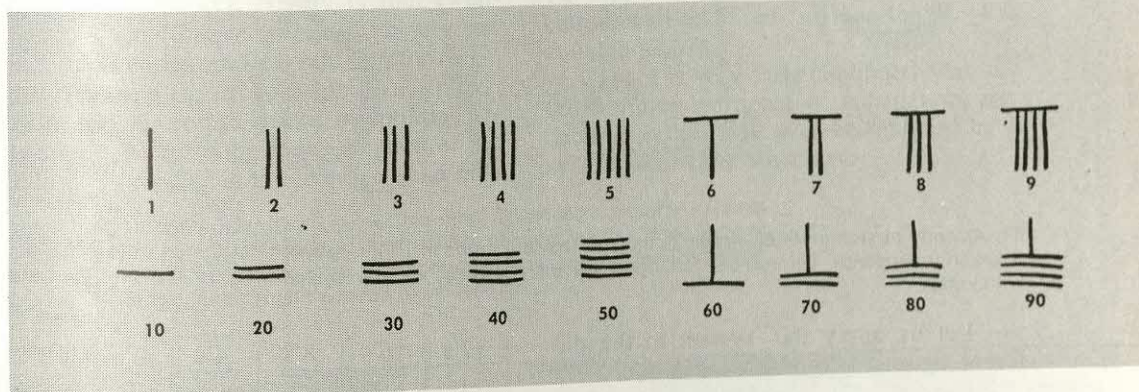
Of course extremely large or extremely small numbers written in this way would be awkward to write and use. We show elsewhere how to indicate such numbers conveniently.

Not all place systems need have ten as their base. Various other bases have been used, or proposed. To show the difference between systems having different bases, let us examine the following paragraph:

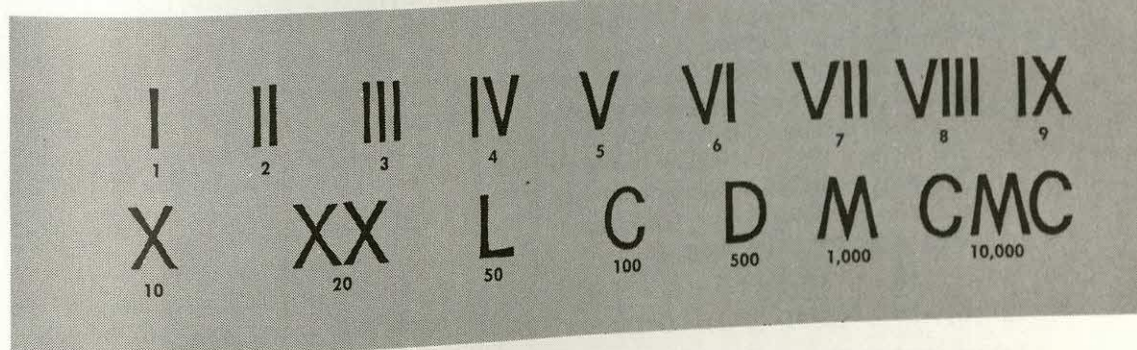
"A boy entered the first grade at the age of 10. After attending school for 22 years, he entered college at the age of 32.

He was 41 years old when he graduated from college, and he married at the age of 100. He died at the age of 240."

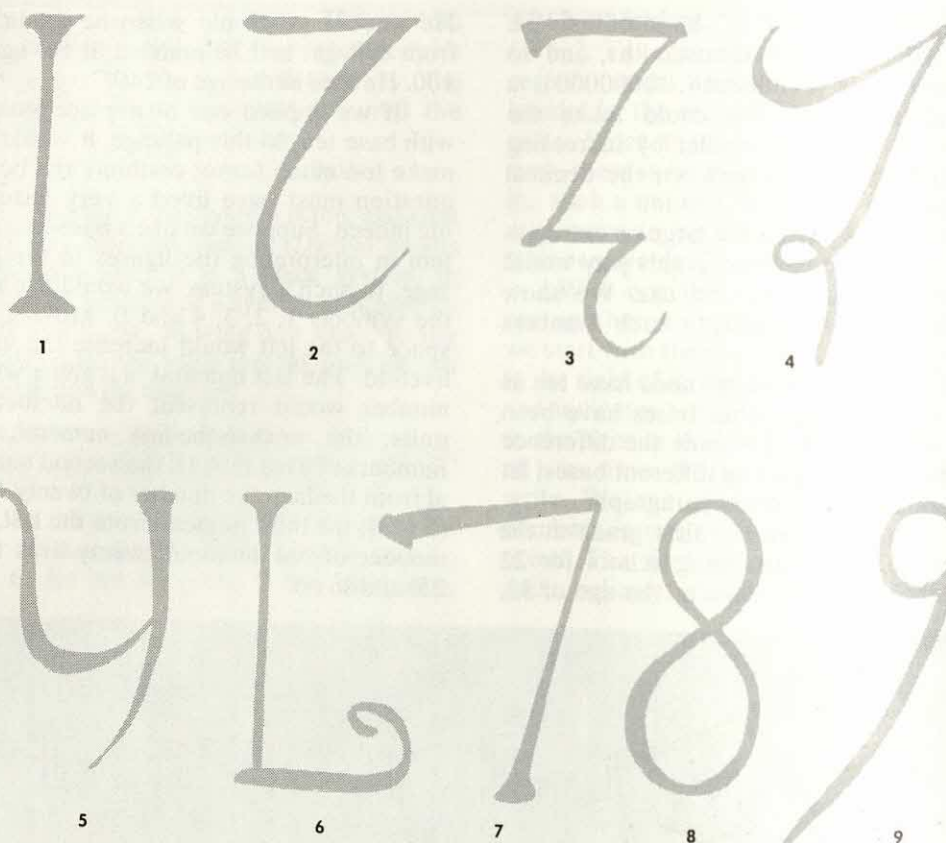
If we applied our own place system, with base ten, to this passage, it would not make too much sense; certainly the boy in question must have lived a very leisurely life indeed. Suppose we use a base-five system in interpreting the figures in the passage. In such a system, we would use only the symbols 1, 2, 3, 4 and 0. Moving one space to the left would increase the value fivefold. The last numeral in a given whole number would represent the number of units; the next-to-the-last numeral, the numbers of fives (5×1); the second numeral from the last, the number of twenty-fives (5×5); the third numeral from the last, the number of one hundred twenty-fives (5×25) and so on.



9. The simple rodlike symbols shown above were employed by the ancient Chinese in their system of writing numbers.



10. The symbols of the Roman numeral system were letters. This system is still used today for certain purposes, such as chapter numbers in some books.



11. An early version of Hindu-Arabic numerals. They are found in a 10th century Spanish manuscript. The symbols changed gradually to those we are familiar with today.

Let us apply this system to the different figures in the above passage. 10 = one five and no units = five; 22 = two fives and two units = twelve; 32 = three fives and two units = seventeen; 41 = four fives and one unit = twenty-one; 100 = one twenty-five, no fives and no units = twenty-five; 240 = two twenty-fives, four fives and no units = seventy. Interpreted in this way, the passage is a perfectly reasonable statement.

The place system called the binary system has the base two. It has only two symbols, 1 and 0. In the binary system, moving one space to the left doubles the value as indicated in the following diagram:

| 32's | 16's | 8's | 4's | 2's | 1's |
|--------|-------|-------|-------|-------|-----|
| (2×16) | (2×8) | (2×4) | (2×2) | (2×1) | |

Let us see how the number 101101 would fit in this scheme:

| 32's | 16's | 8's | 4's | 2's | 1's |
|------|------|-----|-----|-----|-----|
| 1 | 0 | 1 | 1 | 0 | 1 |

As the diagram shows, 101101 in the binary system represents $(1 \times 32) + (0 \times 16) + (1 \times 8) + (1 \times 4) + (0 \times 2) + (1 \times 1) = 32 + 8 + 4 + 1 = 45$.

The binary system was ardently advocated by the great 17th century German philosopher and mathematician Gottfried Wilhelm von Leibniz because of its simplicity and also because he thought that it mirrored creation. Unity (1), he thought, represented God; zero (0) stood for the void from which all things were created.

The binary system has certain practical uses in physics. It also serves in the calculating device called the electronic computer, or "electronic brain." This device is run by electricity; the current is either off or

on. When it is off, the "0" in the binary system is indicated; when the current is on, the number "1" is indicated. The electronic computer can perform the most intricate calculations with the "0" and "1" of the binary system. A complex modern computer is shown in the photo below.

The Mayan Indians, who developed a remarkable civilization in Middle America in the first centuries A.D., used a base-twenty system. There are traces of this method of counting in various modern languages, including French. The French numerals, through sixty-nine, correspond to ours. But seventy in French is sixty-ten (*soixante-dix*); seventy-one is sixty-eleven (*soixante et onze*), and so on until we come to eighty. Eighty is four twenties (*quatre-vingts*). Then come four twenties and one (*quatre-vingt-un*), four twenties and two (*quatre-vingt-deux*). Ninety is four twenties and ten (*quatre-vingt-dix*); ninety-one is four twenties and eleven (*quatre-vingt-onze*) and so on up to a hundred. In English, too, there are traces of counting by twenties. In the King James version of the Bible, we read that the average span of a

human lifetime is "three-score and ten." Abraham Lincoln began his immortal Gettysburg address of November 19, 1863, with the words "Four score and seven years ago."

The Duodecimal Society of America has advocated the adoption of the duodecimal or "base-twelve" system. In this, there are twelve symbols: 1, 2, 3, 4, 5, 6, 7, 8, 9, *t* (standing for ten), *e* (standing for eleven) and 0. Moving one space to the left, in the duodecimal system, increases the value twelvefold, as follows:

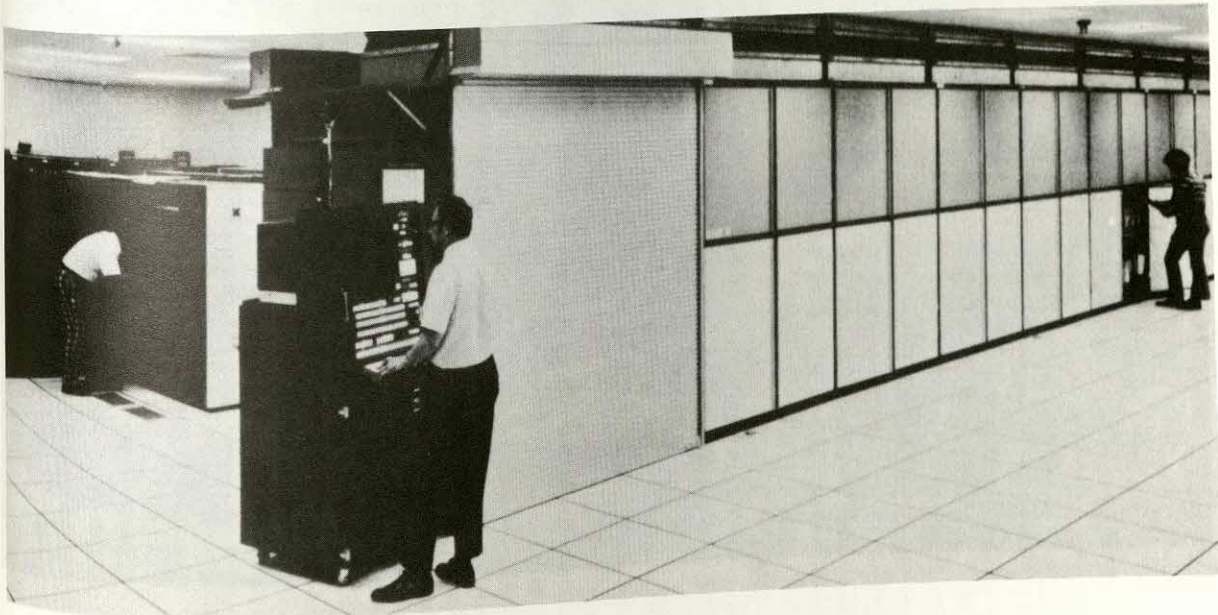
$$\begin{array}{c|c|c|c|c} 20,736's & 1,728's & 144's & 12's & 1's \\ \hline (12 \times 1,728) & (12 \times 144) & (12 \times 12) & (12 \times 1) & \end{array}$$

The number *6et4* in the duodecimal system represents $(6 \times 1,728) + (11 \times 144) + (10 \times 12) + (4 \times 1)$. If we perform these multiplications and additions, we see that *6et4* is equal to 12,076.

There is very little likelihood that, for the ordinary purposes of calculation, our decimal system will ever be replaced by a system with some other base.

The ILLIAC IV computer shown below and other advanced computer systems can process hundreds of millions of instructions per second. The development of computers—an outgrowth of the development of the binary numeral system—has significantly changed many aspects of modern life.

Burroughs



ARITHMETIC

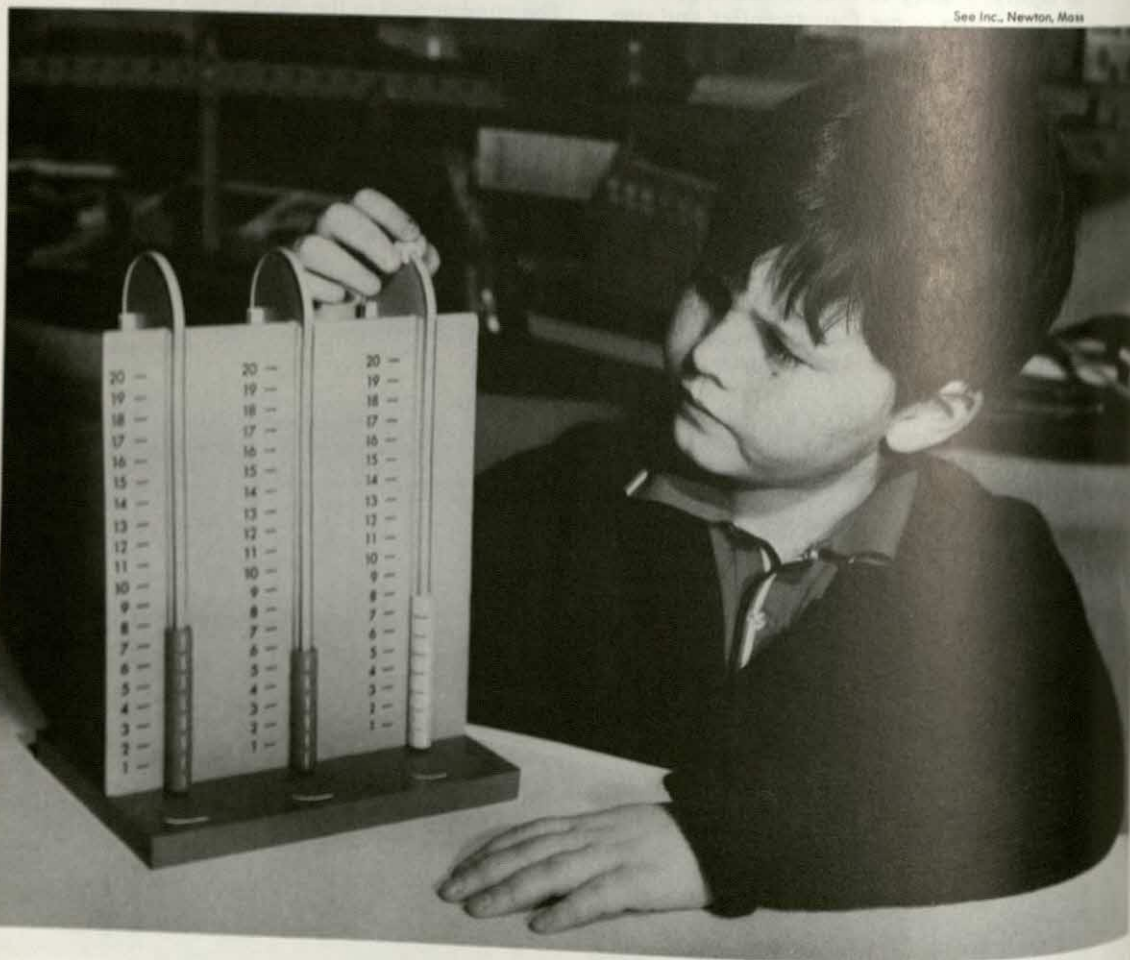
The computations carried out with the numbers of the decimal system make up the branch of mathematics called arithmetic. There are six fundamental operations in arithmetic: addition, subtraction, multiplication, division, involution (obtaining powers of numbers), and evolution (obtaining the roots of numbers).

ADDITION

Addition represents the grouping or bunching together of numbers. If a primitive man wanted to find out how many skins

he would have if he added three skins to two skins, he would lay three skins on the ground, put down two more, and then count the total number. After he had solved this particular problem again and again, he would no longer have to lay out the skins on the ground. He would recall that whenever he added three skins to two skins, the result would always be five skins, and he would do the sum $3 + 2 = 5$ in his head. Similarly, he would come to understand that $1 + 1 = 2$, $1 + 2 = 3$, $1 + 3 = 4$ and so on.

See Inc., Newton, Mass



One can add any sum whatsoever if one can add up to $9 + 9$. Suppose the problem is $23 + 49$. Remember that 23 stands for 2 tens + 3 units and that 49 stands for 4 tens + 9 units. We can state our problem as follows:

$$\begin{array}{r} 2 \text{ tens} + 3 \text{ units} \\ + 4 \text{ tens} + 9 \text{ units} \end{array}$$

First, we would add the units together: $9 + 3 = 12$. 12 is really equal to 1 ten plus 2 units. Keeping the 2 in the units column, we would add the 1 to the other numerals in the tens columns, giving $2 + 4 + 1$ tens, or 7 tens. The answer then is 7 tens + 2 units, or 72.

SUBTRACTION

In subtraction, we take one or more objects from another group of objects. Suppose our primitive mathematician had nine skins and wanted to find out how many skins he would have left if he took five skins away. He would lay out the nine skins on the ground, would take away five skins, and then would count the skins that remained. Later on, he would come to realize that five from nine always leaves four, and he would perform that subtraction in his head. He would then extend this type of calculation to other numbers. He would come to know, for example, that $9 - 4 = 5$ and that $7 - 6 = 1$. He would need to know only a few such calculations to be able to subtract any sum from any other sum.

MULTIPLICATION

Multiplication is really a form of addition. If we did not know how to multiply, we could find the answer to the problem 5×7 by simple addition. For 5×7 means the same thing as five sevens, or $7 + 7 + 7 + 7 + 7$. Adding the five sevens together, we would have 35. Sooner or later we would probably come to realize that when we solve the problem of adding together five sevens, $7 + 7 + 7 + 7 + 7$, the answer is always 35. We would memorize

this particular operation and other similar ones such as three eights (3×8) and seven nines (7×9). These operations up to and including 12×12 are found in the familiar multiplication table. If we could get as far as 9×9 , we could do any multiplication problem, no matter how difficult.

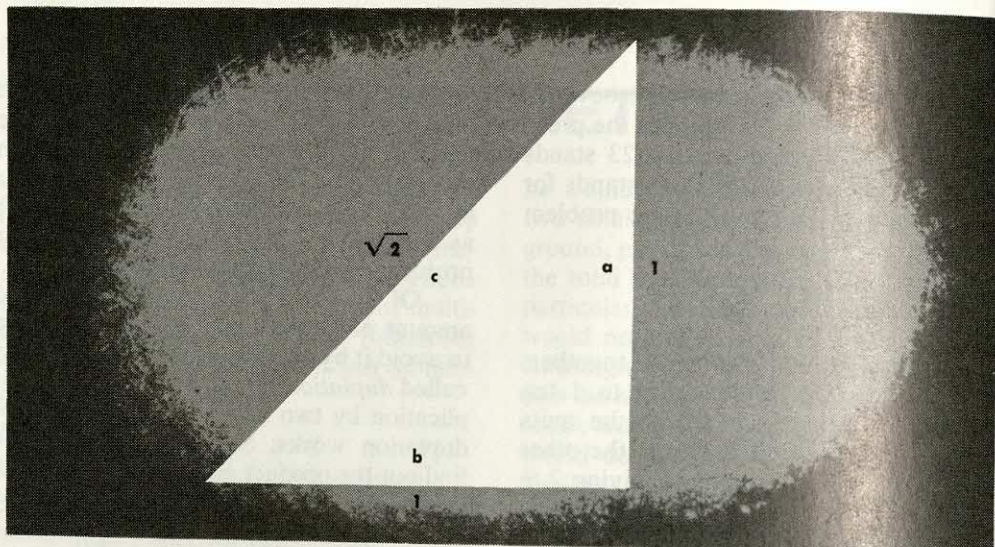
Of course all this requires a certain amount of memorizing. Some people used to avoid it by using a multiplication method called *duplation*, which involves only multiplication by two and addition. This is how duplation works. Suppose we wanted to find out the product of 24 and 18 (24×18). We would perform our calculations thus:

| (a) | (b) | (c) | (d) |
|-----|-----|-----|------------|
| 1 | 24 | | |
| 2 | 48 | 2 | 48 |
| 4 | 96 | | |
| 8 | 192 | | |
| 16 | 384 | 16 | <u>384</u> |
| | | | 432 |

432, in the last column, is the answer to our problem. All this looks quite complicated, but it is really simple.

As you see, we set up four columns, (a), (b), (c), and (d). The first number in column (a) is always 1; the first number in (b) is always the *multiplicand*, which in this case is 24. The multiplicand is a number that is to be multiplied by another number, known as the *multiplier*. We keep on multiplying the numbers in column (a) by 2 until we reach a number that is just equal to or less than the multiplier, which is 18. We stop at 16, since 2 times 16, or 32, is larger than 18. We now double the multiplicand in column (b), giving 48. We double 48, and we keep on doubling the numbers in the column until there are as many numbers in column (b) as in column (a). Next, we select from column (a) the numbers that, added together, would be equal to the multiplier, 18. The numbers we select are 2 and 16. We put these numbers in column (c) in the same places they occupy in column (a). In column (d), we put the column (b) numbers that are next to 2 and 16 in column (a). The sum of the numbers in column (d), 432, is the answer to the problem 24×18 . Of course anyone

A young boy using a loop abacus. The abacus is a device that has been used for thousands of years to perform simple arithmetic calculations.



1. This figure is a right triangle (a triangle having a right angle). Side $a = 1$; side $b = 1$. The square of side c is equal to the sum of the squares of the other two sides. Hence $c^2 = 1 + 1 = 2$. If $c^2 = 2$, $c = \sqrt{2}$.

knowing the multiplication table could multiply 24 by 18 in much less time than it would take to solve the problem by means of duplation.

DIVISION

Division is a kind of subtraction. If we divide 12 by 4, we want to know how many times 4 goes into 12. We perform the following series of subtractions:

$$\begin{array}{r} 12 \\ -4 \\ \hline 8 \end{array} \quad \begin{array}{r} 8 \\ -4 \\ \hline 4 \end{array} \quad \begin{array}{r} 4 \\ -4 \\ \hline 0 \end{array}$$

We have subtracted 4 from 12; then 4 from the remainder 8; then 4 from the remainder 4. Nothing remains. We used 4 as the subtracter three times. Hence the answer to the problem 12 divided by 4, or $12 \div 4$, is 3.

INVOLUTION

In the process called involution, we raise a number to any desired power. The number that is to be raised to the power in question is called the *base*. To raise 2 to the third power, we repeat the base three times as it is multiplied by itself, thus: $2 \times 2 \times 2$. We can indicate the power by placing a small figure, called the exponent to the right of the base and above it. For example, 2 to the third power, or $2 \times 2 \times 2$, is written 2^3 .

If the exponent is 1, it indicates that the number is not raised to a higher power but remains unchanged. Thus $2^1 = 2$; $5^1 = 5$. When we use the exponent zero, we show that the base is to be divided by itself. Any number with the exponent 0 is always equal to 1. For example, $4^0 = 4 \div 4 = 1$; $10^0 = 10 \div 10 = 1$. There are also negative exponents. A base with a negative exponent is equal to the reciprocal of the base $\left(\frac{1}{\text{base}}\right)$ with the corresponding positive exponent. Thus $5^{-1} = \frac{1}{5^1} = \frac{1}{5}$; $3^{-2} = \frac{1}{3^2} = \frac{1}{9}$.

EVOLUTION

In the process called evolution, we find the roots of numbers. Given a certain number, we try to find what other number, multiplied by itself the desired number of times, will give the first number. The number to be multiplied by itself is called the *root*. If we want to find out what number multiplied by itself will give 4, we say that we are trying to find the square root of 4, written $\sqrt{4}$. The number 2, multiplied by itself—that is, used as a factor twice—gives 4. Therefore 2 is the square root of 4. A square root of 4 could also be -2 , since $-2 \times -2 = 4$.

The cube root of the number 8, written as $\sqrt[3]{8}$, is the number which, used as a

factor three times, gives 8. $2 \times 2 \times 2 = 8$; therefore $\sqrt[3]{8} = 2$.

FRACTIONS

Numbers such as 1, 2, 3, 5, 10, 120 and 3,000 are called *whole numbers*, or *integers*. When one multiplies one integer by another, the answer is always an integer: $5 \times 6 = 30$; $7 \times 9 = 63$. However, it is not always possible to obtain an integer as an answer when one divides one integer by another. It is true that if we divide 8 apples into 4 equal shares, each share will consist of 2 apples. But if 8 apples are to be divided into 3 equal shares, the answer will not be an integer. For $8 \div 3 = \frac{8}{3}$, which is a fraction, or "broken number." This does not mean that we have to divide all our 8 apples into thirds and give each person 8 thirds. We note that $\frac{3}{3}$ is exactly equivalent to a

whole apple and that $\frac{6}{3}$ is exactly equivalent to 2 whole apples. To divide 8 apples into 3 equal shares, therefore, we give each person 2 apples. Then we divide the 2 remaining apples into thirds, and given each person 2 of the thirds. Another way of putting it is that $8 \div 3 = 2\frac{2}{3}$. $2\frac{2}{3}$ represents the sum of an integer and a fraction, and is called a *mixed number*.

In a fraction such as $\frac{2}{5}$, the number above the line is called the *numerator*; the number below the line is the *denominator*. The mathematician does not think of $\frac{2}{5}$ as two numbers, but as a single number having a definite value. If we write $\frac{2}{5}$ in the decimal form, 0.4, we can see that the fraction is really a single number.

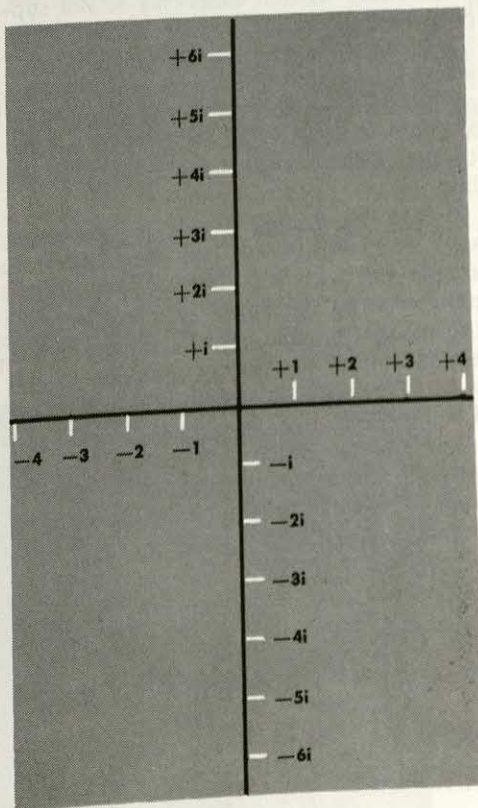
Each fraction is a quotient consisting of two integers. It is possible to write out all the fractions, using the following scheme. First, we set down the fractions with the denominator 1, then those with the denominator 2, then those with the denominator 3, as follows:

Denominator 1 . . . $\frac{1}{1}, \frac{2}{1}, \frac{3}{1}, \frac{4}{1}, \frac{5}{1}$ and so on.
 Denominator 2 . . . $\frac{1}{2}, \frac{2}{2}, \frac{3}{2}, \frac{4}{2}, \frac{5}{2}$ and so on.
 Denominator 3 . . . $\frac{1}{3}, \frac{2}{3}, \frac{3}{3}, \frac{4}{3}, \frac{5}{3}$ and so on.

Obviously, it would take an eternity to write out all the fractions in this way. However, assuming that such a list of fractions could ever be complete, any fraction whatsoever would find a place in it.

Fractions can be added, subtracted, multiplied, and divided according to the usual rules of arithmetic. The answer will always be an integer, or a fraction (consisting of two integers), or a mixed number (consisting of a whole number plus a fraction). In dividing a number by a fraction, we use the rule "Invert the divisor and multiply." Thus $\frac{2}{5} \div \frac{7}{9} = \frac{2}{5} \times \frac{9}{7} = \frac{18}{35}$.

2. In this diagram, the plus Arabic numerals represent units to the east; the minus Arabic numerals, units to the west; the plus *i*'s, units to the north; the minus *i*'s, units to the south. *i* is the symbol for $\sqrt{-1}$.



In a fraction such as $\frac{18}{35}$, the denominator, 35, is really a divisor, since $\frac{18}{35} = 18 \div 35$. Now in mathematics zero is always barred as a divisor, and so it can never be a denominator. This is done to avoid absurd results. For instance, if we were to allow division by zero, we could prove that $6 = 1$. This is how. We know that $6 \times 0 = 0$. If we divide each side of the equation by zero, we have $\frac{6 \times 0}{0} = \frac{0}{0}$, which could be written as $6 \times \frac{0}{0} = \frac{0}{0}$. Now any number divided by itself is equal to 1. Hence we could change the preceding equation to read $6 \times 1 = 1$, or $6 = 1$. In fact, if we were to allow division by zero, we could prove almost anything.

NEGATIVE INTEGERS

If we return now to the integers, we note that when we subtract one positive integer from another, the answer is not always positive. It is true that $7 - 4 = +3$, which is generally written simply as 3. But suppose we want to subtract 7 from 4. To make the subtraction $4 - 7$ possible, we invent a new kind of number called a negative integer, which is represented by an integer with a minus sign in front of it. The answer to the above subtraction, $4 - 7$, would be -3 .

The mechanics of solving problems such as $12 - 6$, $9 - 7$, $6 - 13$ and so on is simple enough. We subtract the smaller number from the larger one and we give the result the sign (plus or minus) of the larger number. By way of example, $12 - 6 = 6$; $9 - 7 = 2$; $4 - 1 = 3$; $4 - 9 = -5$; $3 - 7 = -4$; $6 - 13 = -7$.

There are as many possible negative integers as there are positive integers. Starting with 0, we can build up a list of negative integers to the left of 0 and a list of positive integers to the right of 0, as follows:

$\dots -4, -3, -2, -1, 0, +1, +2, +3, +4 \dots$

We could extend the list of positive and

negative integers in this way indefinitely. There are negative fractions as well as negative integers. Every positive fraction, such as $\frac{3}{5}$, has a negative counterpart, such as $-\frac{3}{5}$.

The entire set of numbers we have been discussing hitherto—positive integers, negative integers, positive fractions, negative fractions, and zero—is called the *rational number system*. We can define a rational number as a zero, or an integer (positive or negative), or a fraction (positive or negative) whose numerator and denominator are both integers. 5 is a rational number; so is -8 ; so is $\frac{1}{2}$; so is $-\frac{3}{4}$.

Rules have been devised for the addition, subtraction, multiplication, and division of positive and negative rational numbers, so that no illogical results will occur. One of these rules is: "The product of two like-signed numbers is positive." For example, $4 \times 4 = 16$; likewise $-4 \times -4 = 16$. Another of these rules is: "The product of two unlike-signed numbers is negative." Examples of the rule are $4 \times -4 = -16$; $-7 \times 6 = -42$.

IRRATIONAL NUMBERS

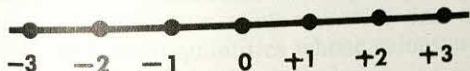
The square root of a number, as we have seen, is one of its two equal factors. Thus the square root of 25, or $\sqrt{25}$, is 5, since $5 \times 5 = 25$. But the square root of 2, or $\sqrt{2}$, cannot be expressed by any rational number. We know that $\sqrt{2}$ must be somewhere between 1.4 and 1.5, since $1.4^2 = 1.96$ and $1.5^2 = 2.25$. We could come closer to the square of 2 by using more and more decimal places. Thus $1.41^2 = 1.9881$ and $1.414^2 = 1.999396$. But even if we used a root with a dozen decimal places, we could never come to a rational number whose square would be 2.

Yet we must deal with numbers such as $\sqrt{2}$ in mathematical work. For example, one of the most familiar theorems in geometry is that in any right triangle (a triangle having a right angle), the square of the hypotenuse (the side opposite the right

angle) is equal to the sum of the squares of the other two sides. Figure 1 represents a right triangle, in which side a is equal to 1 and side b is also equal to 1; c is the hypotenuse. c^2 , therefore, is equal to $a^2 + b^2$, and since $a^2 = 1$ and $b^2 = 1$, $c^2 = 2$. c , therefore, must be equal to $\sqrt{2}$. Obviously, $\sqrt{2}$ is a real number, obtained by a bona fide mathematical calculation, but it is not a rational number. We call it an *irrational number*.

There are a vast number of irrational numbers besides $\sqrt{2}$. The square root of 3 ($\sqrt{3}$) is an irrational number. So is the square root of 5 ($\sqrt{5}$). So is the cube root of 7 ($\sqrt[3]{7}$). These are all real numbers, even if we cannot express them by integers, or by quotients made up of integers.

If we combine all the irrational numbers with all the rational numbers, we get a very large set of numbers called the *real number system*. All the numbers of the system can be represented by points on a straight line. Let one point represent the number 0. Then let points to the right represent positive integers and those to the left, negative integers, as follows:



If the space between the numbers shown above is subdivided into as many parts as possible, each of the subdivision points will represent a rational number. No matter how many divisions we may make in this

way, however, gaps will always remain. If we fill in the gaps with points representing all possible irrational numbers, such as $\sqrt{2}$, $\sqrt{5}$, and so on, the line will be completely filled. This line is known as the real number axis, or *continuum*.

IMAGINARY NUMBERS

There are still other numbers besides the real numbers. Let us consider the number $\sqrt{-1}$. This seems to be a contradiction in terms, since a square is always the product of two equal numbers with like signs and is always positive. Hence no number multiplied by itself can give a negative real number, and it would seem futile to try to get the square root of such a number. However, mathematicians have used the number $\sqrt{-1}$ to form the basis of a number system called *imaginary numbers*, or *complex numbers*. $\sqrt{-1}$ is often indicated by the symbol i .

Complex numbers are used in physics to represent forces. Suppose that we draw two lines intersecting at right angles, as in Figure 2. We use $+1$ to stand for one unit to the east and -1 to stand for one unit in the opposite direction, or the west. We can then interpret $+i$ ($+\sqrt{-1}$) as one unit to the north of the east-west line, and $-i$ ($-\sqrt{-1}$) as one unit to the south. The notation $5 + 3i$ would mean: "Go 5 units to the east and 3 units to the north." $-3 + (-2i)$ would mean: "Go 3 units to the west and 2 units to the south."

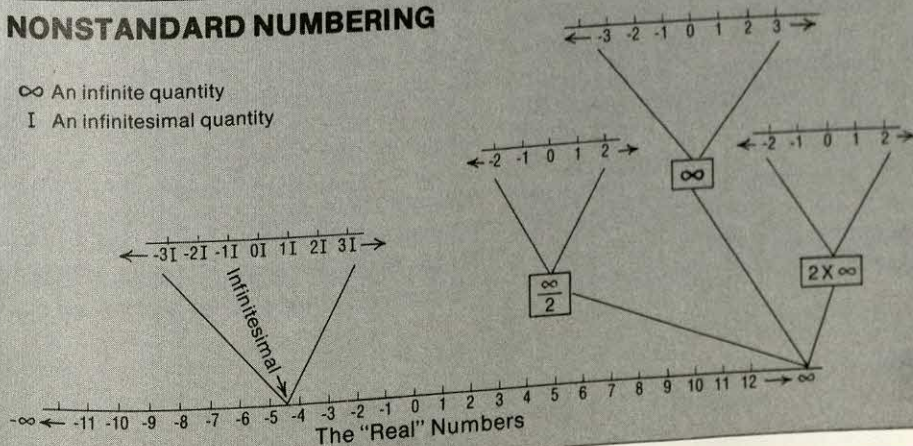
A new system of numbering—one that would be useful in higher mathematics—has been proposed. In this system, known as nonstandard numbering, the interval between ordinary real numbers can be viewed as composed of a series of nonstandard numbers. Nonstandard numbers, viewed from within the interval, can be thought of as ordinary real numbers; from outside the interval, the series of numbers is infinitely small (left) or infinitely large (right).

© 1975 by The New York Times Company, Reprinted by permission.

NONSTANDARD NUMBERING

∞ An infinite quantity

I An infinitesimal quantity



ALGEBRA

by Howard F. Fehr



In the article "Arithmetic," we were concerned with particular numbers, which were expressed by symbols. "Sixty-seven" is a particular number. To do arithmetical problems in which sixty-seven plays a part, we use the symbols "6" and "7," combined as 67. We are now going to consider a branch of mathematics in which a symbol, such as a letter— a , or b , or c —stands not for a particular number, but for a whole class of numbers. This kind of mathematics is called *algebra*.

We can illustrate the difference between arithmetic and algebra by a very simple example. Let us take the number 4. Multiply it by 5 ($4 \times 5 = 20$); add 4 ($20 + 4 = 24$); double the answer ($24 \times 2 = 48$); subtract 8 ($48 - 8 = 40$); divide by the original number, 4 ($40 \div 4 = 10$). The result of all these operations, as you see, is 10. In arriving at the final result, 10, we used the method of arithmetic, involving particular numbers, throughout.

Suppose now that we think of any

number. Let us indicate "any number" by the symbol x and let us go through the same operations as before. We multiply x by 5 ($5 \times x$, written $5x$); add 4 ($5x + 4$); double the answer ($10x + 8$); subtract 8 ($10x$); divide by the original number, which is x ($\frac{10x}{x} = 10$). The answer, then, is 10.

Here we have been using the methods of algebra, because x can be replaced by any number. We could substitute for it 2, or 3, or 15, and the final result would be 10.

When a generalized number, represented by a letter (such as a) is multiplied by a particular number (such as 5) or by another generalized number (such as b), we do not use multiplication signs, but indicate multiplication by putting these symbols close to one another. Thus $a \times b = ab$; $5 \times a = 5a$; $5 \times a \times b = 5ab$. We could not indicate the multiplication of two particular numbers in this way. 7×5 could not be given as 75, because 75 really stands for $70 + 5$.

Let us consider another example. In the equation $(2 + 3)^2 = 25$, we are dealing with the particular numbers 2 and 3, and the result is always 25. But suppose that instead of two particular numbers, we used the letter a , standing for any number, and the letter b , standing for any number other than a . We would then have $(a + b)^2 = a^2 + 2ab + b^2$.

What is significant about $(a + b)^2 = a^2 + 2ab + b^2$ is that it indicates a general relationship that holds true for a great many particular numbers. If we substituted 3 for a and 2 for b , we would have $(3 + 2)^2 = 3^2 + (2 \times 3 \times 2) + 2^2 = 9 + 12 + 4 = 25$. Or we could substitute 5 for a and 6 for b , giving $(5 + 6)^2 = 5^2 + (2 \times 5 \times 6) + 6^2 = 25 + 60 + 36 = 121$.

Algebra, the mathematics of "any numbers," or "variables," goes to the heart of the relationship between numbers. Generally speaking, it is concerned with particular numbers only insofar as they are applications of general principles. It is also used in the solution of certain specific problems in which we start out with one or more unknown quantities whose values are indicated by algebraic symbols.

AN ANCIENT DISCIPLINE

The study of algebra goes back to antiquity. Recent discoveries have shown that the Babylonians solved problems in algebra, although they had no symbols for variables. They used only words to indicate such numbers, and for that reason their algebra has been referred to as rhetorical algebra. The Ahmes Papyrus, an Egyptian scroll going back to 1600 B.C., has a number of problems in algebra, in which the unknown is referred to as a *hau*, meaning "a heap."

Little further progress was made in algebra until we come to Diophantus, a Greek mathematician, living in the third century A.D. He reduced problems to equations, representing the unknown quantity by a symbol suggesting the Greek letter Σ (sigma). He also introduced an interesting system of abbreviations, in which he used only the initial letters of words and omitted all unnecessary words. If we were to use the method of Diophantus in present-

ing the problem: "An unknown squared minus the unknown will give twenty," we would first state the problem as "Unknown squared minus unknown equals twenty." Then we would use initial letters for all the words except the last and we would give the numeral for "twenty," as follows: "USMUE20."

In the sixteenth century, François Vieta, a French mathematician, used the vowels a, e, i, o, u to represent unknown numbers and the consonants b, c, d, f, g , and so on to stand for values that remained fixed throughout a given problem. The great 17th-century French philosopher René Descartes proposed the system of algebraic symbols now in use. In this system, a, b, c and other letters near the beginning of the alphabet represent the fixed numbers. The last letters of the alphabet — x, y, z and also sometimes w — stand for the unknown numbers in a problem. As soon as this symbolism came into general use, algebra grew quite rapidly into a systematic set of rules and theorems that could be applied to all numbers.

The word "algebra" originated from the title of a work on algebra by a Persian, Mohammed ibn Musa Al-Kwarizmi, who lived in the ninth century A.D. He wrote in Arabic a work called *Al-Jabr W'al Muqabala*, which means "restoration and reduction." By *al-jabr*, or restoration, was meant the transposing of negative terms to the other side of an equation to make them positive. When the Arabs came to Spain, they brought this word with them. In the course of time, *al-jabr* was changed to "algebra," and the word came to be applied not to a single operation, but to the many operations involved in algebra.

THREE FUNDAMENTAL LAWS

Algebra generalizes — that is, expresses in general terms — certain basic laws which govern the addition, subtraction, multiplication, and division of all numbers.

(1) When we add or multiply two integers, the order in which we add or multiply them is immaterial. Thus $2 + 3$ is the same as $3 + 2$, and 4×3 is the same as 3×4 . Since this is true for all integers,

we set up the following algebraic formulas:

$$a + b = b + a; ab = ba$$

These are called the *commutative laws* of addition and multiplication.

(2) When more than two numbers are added or multiplied, we can group them in any order we choose, and the answer will always be the same. If 2 is added to (3 + 6), the result is the same as if we added (2 + 3) to 6. Similarly, 2 times the product of 3×6 is the same as 3 times the product of 6×2 . These results are indicated in the formulas:

$$a + (b + c) = (a + b) + c; a(bc) = (ab)c$$

These are the *associative laws* of addition and multiplication.

(3) If a multiplicand has two or more terms, a multiplier must operate upon each of these terms in turn. Suppose we wish to multiply $3 + 2$ by 5, a problem which we could set down as $5(3 + 2)$. We would first multiply 3 by 5 and then 2 by 5, giving $15 + 10$, or 25. This rule is called the *distributive law of multiplication* and is given by the following formula:

$$a(b + c) = ab + ac$$

Suppose we want to multiply $a + b$ by $a + b$, which of course would be the same thing as $(a + b)^2$. We would set up the problem as follows:

$$\begin{array}{r} a + b \\ a + b \end{array}$$

In accordance with the distributive law of multiplication, we first multiply a and b on the upper line by b in the lower line. Then we multiply a and b on the upper line by a of the lower line. Finally, we add the results:

$$\begin{array}{r} a + b \\ a + b \\ \hline ab + b^2 \\ a^2 + ab \\ \hline a^2 + 2ab + b^2 \end{array}$$

We use the same distributive law of multiplication in multiplying 25 by 25. Ordinarily, we would present our calculations as follows:

$$\begin{array}{r} 25 \\ 25 \\ \hline 125 \\ 50 \\ \hline 625 \end{array}$$

Actually, since the 2 in 25 is really 20, the problem is

$$\begin{array}{r} 20 + 5 \\ 20 + 5 \end{array}$$

Using the distributive law, we multiply 5 in the first line by 5 in the second line and 20 in the first line by 5 in the second line. Then we multiply 5 in the first line by 20 in the second line and 20 in the first line by 20 in the second line. Finally we add the products. We could indicate these operations as follows:

$$\begin{array}{r} 20 + 5 \\ 20 + 5 \\ \hline 100 + 25 \\ 400 + 100 \\ \hline 400 + 200 + 25 = 625 \end{array}$$

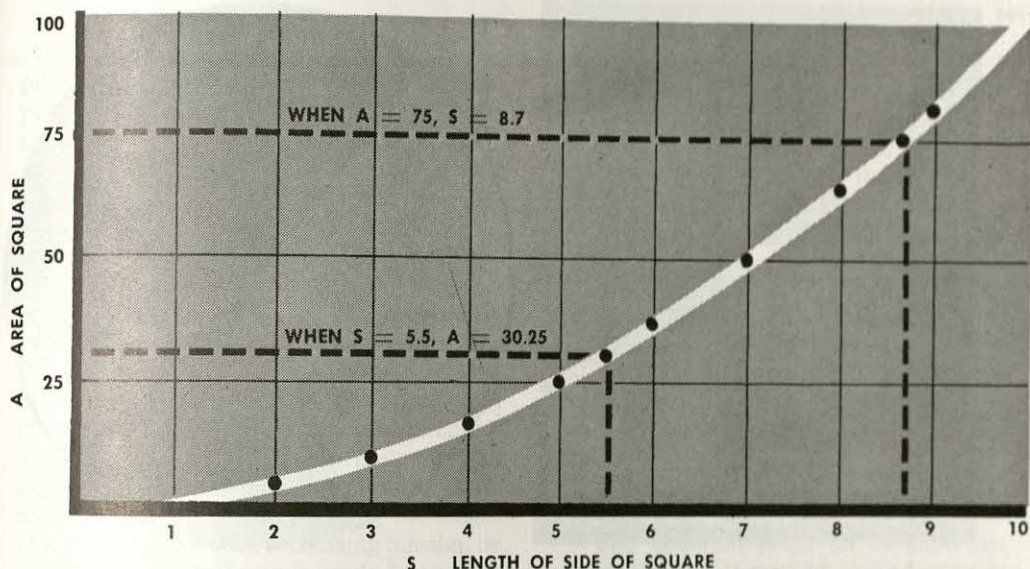
If you carefully compare (I) and (II) above, you will see that basically the same operations are performed in both. Simply bear in mind that, for example, when in (I) we multiply the 2 in the second line by 5 in the first line, the 2 in the second line stands for 20.

FORMULAS, TABLES, AND GRAPHS

There are different ways of showing how different quantities are related. We can use a formula, set up a table of values, or draw up a graph.

Take the rule: "The area of a square is equal to the square of the length of its side." This is a rather roundabout way of expressing the relationship in question. We could state it much more simply by using the symbol A to represent the number of square units in the area, and the symbol s to represent the number of units in the side. The above rule can then be stated as $A = s^2$. This abbreviated rule is called a formula, from the Latin word meaning "little form."

If the length of the side is 6, we can get A , the area, by substituting 6 for s in



1. Drawing of the graph of the area of a square, using horizontal and vertical axes, according to the method described in the text. The graph shows how the area of a square increases as the length of the side increases. It is one way of expressing the relation between the area and the length of the side.

the formula. A then becomes 6^2 or 36. As you see, we have put our formula to practical use. Technically speaking, we have evaluated the formula for the value 6.

We can also represent the relationship $A = s^2$ by setting up a table of values. Suppose the side of the square is equal to 1; A is then 1^2 or 1. If the side of the square is equal to 2, $A = 2^2$, or 4. Substituting for s the values from 1 through 10 in turn, we obtain the following table of values:

| | | | | | | | | | | |
|-----|---|---|---|----|----|----|----|----|----|-----|
| s | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| A | 1 | 4 | 9 | 16 | 25 | 36 | 49 | 64 | 81 | 100 |

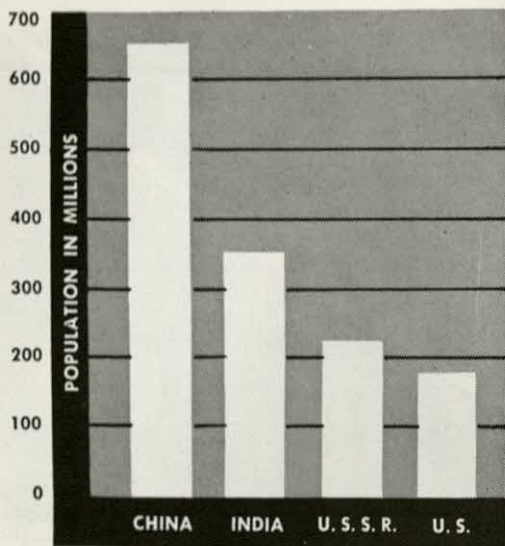
This table tells us that if $s = 1$, $A = 1$; if $s = 2$, $A = 4$; if $s = 3$, $A = 9$, and so on.

There is still another way of representing the area of a square. We could construct a graph, as in Figure 1. First, we draw two lines, called axes, which are perpendicular to each other. Along the horizontal axis, or base, we mark out a series of numbers at equal intervals, corresponding to the values of s in the table. We mark out another series of numbers on the vertical axis. These numbers correspond to the values of A . At each of the values of s , 1, 2, 3, 4, and so on, we erect a perpendicular. Then we mark the appropriate value of A , corresponding to a given value

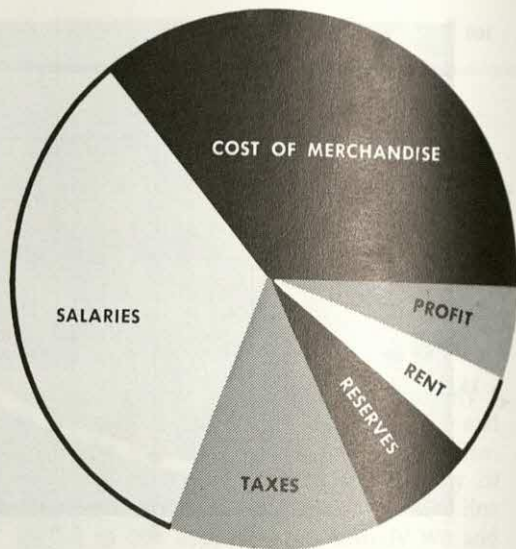
of s , on the vertical scale. We draw a perpendicular at this point, and mark a point where the two perpendiculars meet. When we have performed the same operation for each of the values of s and A , we shall have the series of points indicated in Figure 1. A smooth, curved line is then drawn, connecting the different points. This line is called the graph of the area of the square. It shows how the area increases as the length of the side increases.

We can use the graph to find out the area of squares with sides not given in the tables of values. For example, if $s = 5.5$, we erect a perpendicular at this point, extending it until it meets the graph. From the point of meeting, we erect a perpendicular to the vertical axis. The point where this perpendicular meets the vertical axis will represent the area, which is about 30. The correct value is 30.25.

If we know the area of a square, we can find out the approximate length of the side by means of our graph. Suppose the area is 75. From a point corresponding to the number 75 on the vertical axis, we draw a line parallel to the horizontal axis. From the point where this line meets the curve, we drop a perpendicular to the horizontal axis. This line will meet the horizontal



2. Bar graphs showing the populations of four countries in the mid-20th century.



3. A circle graph indicating how each part of a dollar received by a department store is spent.

axis at about 8.7, the length of the side of the square. Here the correct value, expressed to five decimal places, is 8.66025.

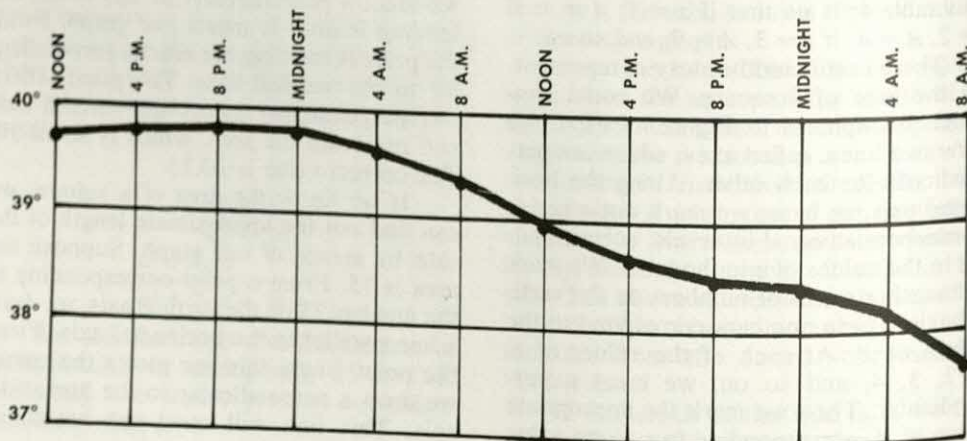
There are many different kinds of graphs. The most common are the bar graph, the circle graph, and the line graph. If, for example, we would like to show how the populations of China, India, the Soviet Union, and the United States compare, we would draw four bars (Figure 2). The lengths of the bars would be proportionate

to the populations in question.

If we desired to compare parts of a whole quantity, we would use a circle graph. It could serve to indicate how each part of a dollar received in sales in a department store is spent by the store (Figure 3). Each sector of the circle, as compared to the whole circle, would show the proportion given to a particular service.

When a quantity is continuously changing, we would use a line graph. In a hos-

4. Nurses sometimes make line graphs of patient's temperatures, taking readings at regular intervals. Readings, in degrees Celsius, taken every four hours would yield a graph like this one.



pital, the nurse makes a graph of the temperature of each patient (Figure 4). She takes the temperature every four hours, say, locates the temperature reading at a given point, and then connects the point with the preceding point by a straight line. The doctor who consults the chart, can see at a glance how the patient's temperature is changing.

To repeat, then, we can show how quantities are related by a formula, a table of values, or a graph. It is the formula that is basic. If a table of values is worked up, a scientist tries to find the formula that will express the relationship in question. If an engineer plots a graph showing the length of a steel cable under increasing tension, he works out an algebraic formula to sum up his findings.

Formulas are extremely important in many branches of pure and applied science. For example, to indicate the speed of a body having uniform rectilinear motion—that is, moving in a straight line at a constant speed—the physicist uses the formula $v = \frac{s}{t}$, where v is the average speed of the body, s is the space or distance covered, and t is the time required to travel this distance. We can apply this formula to specific cases by substituting appropriate values for v , s , or t . For example, if a car takes 5 hours to travel 400 kilometers, as indicated by the odometer, we could find the average speed by substituting 5 for t and 400 for s . The average speed, then, would be $\frac{400}{5}$ kilometers, or 80 kilometers, per hour.

The chemist often has occasion to use the laws of Charles and Gay-Lussac, which state that if the pressure and the mass of a gas are constant, the volume is proportional to the absolute temperature. Absolute temperature is based on the lowest possible temperature, "absolute zero." Absolute zero equals -273.16° Celsius. This law can be stated very concisely by the formula $\frac{V_1}{T_1} = \frac{V_2}{T_2}$, in which V_1 is the volume of a gas at temperature T_1 , while V_2

is the volume of a gas at temperature T_2 .

Perhaps the most famous formula of all is the Einstein Equation, $E = mc^2$. In this equation E stands for the amount of energy, m for the amount of mass, and c for the speed of light (which is about 300,000 kilometers per second), measured in appropriate units. With this formula the great 20th-century physicist Albert Einstein indicated the amount of energy that appears when matter is transformed into energy. The use of atomic energy would not have been possible without this formula.

HOW EQUATIONS ARE EMPLOYED

The equation plays an all-important part in algebra. It may be looked upon as a balance, with equal numerical values on each side of the "equal" sign ($=$). To show how an equation is applied, let us consider the formula for the perimeter (the outside boundary) of a rectangle: $P = 2l + 2w$, where P is the perimeter, l is the length and w is the width. Suppose we have 100 meters of wire with which to make a rectangular enclosure, which is to be 20 meters wide. We wish to find out the length of this enclosure. The formula for the perimeter, we saw, is $P = 2l + 2w$. We know that the perimeter is 100 (the 100 meters of wire at our disposal). We also know that the width is to be 20 meters. Substituting 100 for P and 20 for w in the formula, we have $100 = 2l + 2 \times 20 = 2l + 40$. In the equation $100 = 2l + 40$, the 100 on one side of the equal sign is exactly equal to $2l + 40$ on the other side. To find out what l is, we subtract 40 from each side of the equation. That gives $60 = 2l$. l , therefore, is 30. The length of the enclosure is 30 meters.

In solving the equation, we used the rule: "If the same operation is performed on each member of an equality, then the results are equal." If we add the same quantity to each side of the equation or subtract the same quantity, the equality will be maintained. It will be maintained, too, if we multiply or divide each side of the equation by the same quantity. Of course, if we multiplied each side of the equation by zero, the result would be $0 = 0$, which

would get us nowhere in the task of solving the equation. Division by zero is not permitted.

Equations may be used to solve problems in which no formula is involved but in which certain data are given. Here is a simple problem: "A man is 6 times as old as his son. In 20 years, the father will be only twice as old as his son. How old are the father and son at the present time?"

On the basis of the data, we can write an equation and solve the problem. First, we let x stand for the son's age. Since the father is 6 times as old as his son, his age can be given as 6 times x , or $6x$. In 20 years—that is, when 20 years are added to the son's age—the son will be $x + 20$ years old. In 20 years, the age of the father will be $6x + 20$ years. At that time, the father's age will be twice that of the son, a relationship which we can express by the equation: $6x + 20 = 2(x + 20)$. Applying the distributive law to the right-hand side of the equation, we have $6x + 20 = 2x + 40$. We subtract $2x + 20$ from each side of the equation and get $4x = 20$. If $4x = 20$, $x = 5$. The son's age at the present time, therefore, is 5. Since the father's age at the present time is 6 times that of the son, or $6x$, the father is 30 years old.

Not all problems are as simple as this one, in which the unknown is x . The equation may involve not only an unknown quantity, x , but also higher powers of x . If x^2 is the highest power occurring in an equation, it is called a *quadratic equation*. $x^2 + 6 = 5x$ is an example of such an equation. In various equations, the highest power of x may be x^3 or x^4 , or there may be even higher powers.

IDENTITIES

When an equation is true for all the replacement values of the variables concerned, it is called an *identity*. A familiar example of an identity is $(a + b)^2 = a^2 + 2ab + b^2$. As we pointed out before, this equation holds true no matter what values we assign to a and b . It can be used as an aid in mental arithmetic. To square 22, we can think of this number as $(20 + 2)^2$, 20 being substituted for a and 2 for b in the

above identity. Mentally we square 20, giving 400; then we double 20×2 , giving 80; finally, we square 2, giving 4. $400 + 80 + 4 = 484$. The answer, then, is 484.

Another identity is $(a - b)^2 = a^2 - 2ab + b^2$. You can verify this by performing the multiplication $(a - b)(a - b)$. We would have:

$$\begin{array}{r} a - b \\ - (a - b) \\ \hline a^2 - ab \\ a^2 - 2ab + b^2 \end{array}$$

Still another identity is $(a + b)(a - b) = a^2 - b^2$. This also is useful in certain mental-arithmetic problems. If we wish to multiply 34 by 26 in our heads, we can change the problem to $(30 + 4) \times (30 - 4)$. Solving this in accordance with the identity $(a + b)(a - b) = a^2 - b^2$, we have $900 - 16 = 884$.

Other well-known identities are:

$$(a + b)^3 = (a^3 + 3a^2b + 3ab^2 + b^3)$$

$$(a - b)^3 = (a^3 - 3a^2b + 3ab^2 - b^3)$$

$$(a^3 - b^3) = (a - b)(a^2 + ab + b^2)$$

EXPONENTS

Exponents simplify the writing of algebraic expressions. Thus $aaabbbcc$, which is really a continuous multiplication (a times a times a times b times b times b times c times c), can be written $a^3b^3c^2$. The mathematician has derived a series of rules for combining exponents. The rules are stated in general terms; a^n stands for the *base* a raised to the n th *power*, a^m , for the same base raised to the m th power.

(1) $a^n a^m = a^{n+m}$. In multiplying powers, we add the exponents of like bases. Thus $2^2 \times 2^3 = 2^{2+3} = 2^5 = 32$.

(2) $a^n \div a^m = a^{n-m}$. In dividing powers, we subtract the exponents of like bases. This means that $2^5 \div 2^2 = 2^{5-2} = 2^3 = 8$.

(3) $(a^n)^m = a^{nm}$. To raise a given power by another power, we multiply the two exponents. For example, $(2^2)^3 = 2^{2 \times 3} = 2^6 = 64$.

(4) $(ab)^n = a^n b^n$. When a product is

raised to a power, each member of the product is raised to that power. Thus $(4 \times 2)^2 = 4^2 \times 2^2 = 16 \times 4 = 64$.

(5) $\left(\frac{a}{b}\right)^n = \frac{a^n}{b^n}$. When a quotient is raised to a given power, each member of the quotient must be raised to that power.

$$\left(\frac{2}{3}\right)^3 = \frac{2^3}{3^3} = \frac{8}{27}.$$

It should be noted here that a base with the exponent zero is equivalent to 1. Thus $10^0 = 1$; $3^0 = 1$; $1^0 = 1$. The above rules for exponents apply to zero exponents. For example, $a^na^0 = a^{n+0} = a^n$; $5^3 \times 5^0 = 5^{3+0} = 5^3 = 125$.

A base with a negative exponent is equal to the reciprocal of the base $\left(\frac{1}{\text{base}}\right)$ with the corresponding positive exponent.

Thus $2^{-2} = \frac{1}{2^2}$. Negative exponents follow

the rules for exponents. Thus $a^{-5}a^3 = a^{-5+3} = a^{-2} = \frac{1}{a^2}$; $10^{-7} \times 10^5 = 10^{-7+5} =$

$$10^{-2} = \frac{1}{10^2} = \frac{1}{100}.$$

Exponents can also occur in the form of fractions. Thus we have $a^{1/2}$, $a^{1/3}$, $a^{1/4}$, and so on. $a^{1/2}$ means the square root of a (\sqrt{a}); $a^{1/3}$ means the cube root of a ($\sqrt[3]{a}$); $a^{1/4}$ means the fourth root of a ($\sqrt[4]{a}$). The numerator in fractional exponents need not necessarily be 1. We frequently deal with exponents such as $\frac{2}{3}$ and $\frac{3}{5}$. In such cases,

the numerator stands for a power of a base and the denominator for the root of a base. $10^{2/3}$, for example, is equal to $\sqrt[3]{10^2}$.

All fractional exponents, whether or not the numerator is 1, follow the rule for exponents. For example, $10^2 \times 10^{2/3} = 10^{2+2/3} = 10^{8/3} = \sqrt[3]{10^8}$.

EXPRESSING VERY LARGE OR VERY SMALL NUMBERS

Exponents provide a convenient way of writing very large or very small numbers. We know that 1,000,000 is 10^6 , the exponent 6 representing the number of

zeros after 1. We could indicate 5,000,000 as $5 \times 1,000,000$, or 5×10^6 . To write 5,270,000, we would multiply 1,000,000 or 10^6 by 5.27. The number would be written as 5.27×10^6 . In other words, we can express a large number as the product of two numbers: the first a number between 1 and 10; the second, a power of 10.

The number 5,270,000 is not too formidable and we can grasp it readily enough. But consider the problems that would arise if, in our calculations, we had to use a number such as 602,000,000,000,000,000,000. It represents the number of molecules in 18 grams of water, and it is called Avogadro's number, after the early 19th-century Italian scientist Amedeo Avogadro, who worked out the value. It is used in a great many scientific calculations, but practically never in the form in which we have given it. Instead, it is written as 6.02×10^{23} .

The 20th-century American mathematician Edward Kasner invented a new system of indicating extremely large numbers. He coined the word "googol" to express the number 10^{100} , which would be equivalent to 1 followed by 100 zeros. He invented another term, the "googolplex," to stand for $10^{100^{100}}$, or the figure 1 followed by a googol of zeros—that is, 10,000 zeros.

Exponents can be used just as effectively to express very small numbers. Since a minus exponent indicates how many times

the fraction $\frac{1}{\text{base}}$ is repeated as it is multi-

plied by itself, $10^{-3} = \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} = .001$.

Note that the exponent 3, in 10^{-3} , represents the number of digits after the decimal point in the number .001. .005 could be written as $5 \times .001$, or 5×10^{-3} . Now consider a much smaller number. The wave length of red light is 0.00000077 meters. We can write this number as $7.7 \times .0000001$ meters, or 7.7×10^{-7} meters.

Writing a number as the product of (1) a number between 1 and 10 and (2) a power of 10 is called *scientific notation*. It is widely used by scientists and engineers.

CALCULATIONS WITH LOGARITHMS

Exponents have also been put to work to simplify arithmetical calculations. Suppose that we represent numbers as powers of 2. We know that $2^{-2} = \frac{1}{2^2} = \frac{1}{4}$; $2^{-1} = \frac{1}{2}$; $2^0 = 1$; $2^1 = 2$; $2^2 = 4$; $2^3 = 8$; $2^4 = 16$; and so on. Expressed as a power of 2, therefore, $\frac{1}{4}$, or .25, is 2^{-2} ; $\frac{1}{2}$, or .5, is 2^{-1} ; 1 is 2^0 ; 2 is 2^1 ; 4 is 2^2 ; 8 is 2^3 ; 16 is 2^4 . Let us now make a table setting down (1) certain numbers; (2) these numbers expressed as powers of 2; (3) the exponents in question.

Consider the problem $.25 \times 1024$. The table shows that $.25 = 2^{-2}$ and that $1024 = 2^{10}$. The problem then becomes $2^{-2} \times 2^{10}$. Applying the first law of exponents given above, we have $2^{-2} \times 2^{10} = 2^{-2+10} = 2^8$. Consulting the table, we find that $2^8 = 256$. 256, then, is the answer to the problem $.25 \times 1024$. We have changed a problem in multiplication into a problem in addition—a simpler operation.

Let us take another problem: $1024 \div 32$. Looking at the table, we see that 1024 is 2^{10} and that 32 is 2^5 . Applying the second law of exponents, we have $2^{10} \div 2^5 =$

$2^{10-5} = 2^5$. We now consult the table and find that 2^5 is equal to 32. This is the answer to $1024 \div 32$. In this case, we have changed a problem in division into a simple problem in subtraction.

Our next problem is to raise 4 to the fifth power. In other words, we want to know what 4^5 would be. The table shows us that 4 is 2^2 . From the third law of exponents, we know that $(2^2)^5 = 2^{2 \times 5} = 2^{10}$. 2^{10} , according to the table, is 1024, which is the answer. We have solved our problem by a single multiplication instead of multiplying $4 \times 4 \times 4 \times 4 \times 4$.

Suppose we wish to get the square root of 1024. According to the table, 1024 is 2^{10} . To get the square root of a given power, we divide the exponent indicating that power by 2. Hence the square root of $2^{10} = 2^{10 \div 2} = 2^5$. Dividing the exponent indicating the power by 2 is really in accordance with the third law of exponents. The square root of a number, as we have seen, is equivalent to the same number with

the exponent $\frac{1}{2}$. The square root of 2^{10} , therefore, can be expressed as $(2^{10})^{1/2}$. Remember that to multiply a number by $\frac{1}{2}$ is the same thing as to divide it by 2. The

table shows that $2^5 = 32$. 32, then, is the square root of 1024.

When a number is expressed as a power of a given base—in this case the base two—we call the exponent that indicates the power the *logarithm* of the number. All the exponents in the third column of the table are the logarithms, to the base two, of the numbers in the first column. -2 is the logarithm of .25 when the base is two. As a mathematician would put it, $\log_2 .25 = -2$. Also when the base is two, the logarithm of 4 is 2; the logarithm of 64 is 6. To multiply numbers, we add their logarithms. To divide numbers, we subtract their logarithms. To raise a number to a given power, we multiply the logarithm of the number by the power in question. To obtain the root of a number, we divide the logarithm of the number by the desired root. After we have added, or subtracted, or multiplied, or di-

| NUMBER | NUMBER EXPRESSED AS POWER OF 2 | EXPONENT IN PRECEDING COLUMN |
|--------|--------------------------------|------------------------------|
| .25 | 2^{-2} | -2 |
| .5 | 2^{-1} | -1 |
| 1 | 2^0 | 0 |
| 2 | 2^1 | 1 |
| 4 | 2^2 | 2 |
| 8 | 2^3 | 3 |
| 16 | 2^4 | 4 |
| 32 | 2^5 | 5 |
| 64 | 2^6 | 6 |
| 128 | 2^7 | 7 |
| 256 | 2^8 | 8 |
| 512 | 2^9 | 9 |
| 1024 | 2^{10} | 10 |

| NUMBER | NUMBER EXPRESSED AS POWER OF 10 | LOGARITHM TO THE BASE TEN (\log_{10}) |
|--------|---------------------------------------|---|
| .0001 | 10^{-4} | -4 |
| .001 | 10^{-3} | -3 |
| .01 | 10^{-2} | -2 |
| .1 | 10^{-1} | -1 |
| 1 | 10^0 | 0 |
| 10 | 10^1 | 1 |
| 100 | 10^2 | 2 |
| 1000 | 10^3 | 3 |
| 10000 | 10^4 | 4 |

vided in this way, we find the number that corresponds to the resulting logarithm.

All the logarithms we gave above were to the base two. Most tables of logarithms are given to the base ten. Let us now prepare another table, giving (1) a series of numbers; (2) the numbers expressed as powers of 10; (3) the logarithms of the numbers—that is, the exponents when the numbers are expressed as powers of 10.

To solve the problem $.0001 \times 100$, we consult the table and find the logarithms of .0001 and 100 (-4 and 2, respectively), add the logarithms ($-4 + 2 = -2$) and find the number corresponding to the logarithm -2. This number, as we see from the table, is .01. We can also do such problems as $10,000 \div .0001$, 10^4 and $\sqrt{10,000}$.

Of course, to be serviceable, a table of logarithms would have to include the logarithms of other numbers besides those given above. It would have to give, for example, not only the logarithms of 1 and 10, but also those of 2, 3, 4, 5, 6, 7, 8, and 9. We know that since $1 = 10^0$ and $10 = 10^1$, the logarithm of 2 would be between 0 and 1. Mathematicians have calculated that it is 0.301. This means that, expressed as a power of 10, the number 2 is $10^{0.301}$. The integer part of the logarithm (0 in this case) is called the *characteristic*. The decimal part (.301) is called the *mantissa*.

| NUMBER | LOGARITHM (BASE TEN) |
|--------|-------------------------|
| 1 | 0. |
| 2 | 0.301 |
| 3 | 0.477 |
| 4 | 0.602 |
| 5 | 0.699 |
| 6 | 0.778 |
| 7 | 0.845 |
| 8 | 0.903 |
| 9 | 0.954 |
| 10 | 1. |

We give only three decimal places for the sake of simplicity. Logarithms have been calculated to more than twenty places. Depending upon the accuracy desired, one would use a four-place table, or a five-place table, or a seven-place table and so on.

The logarithms of the other numbers from 3 through 10 have also been worked out, as you can see on the table.

Using the table, let us multiply 2 by 4. We add 0.301, the logarithm of 2, and 0.602, the logarithm of 4, and we get the logarithm 0.903. Consulting the table, we see that 0.903 is the logarithm of 8. A mathematician would say that 8 is the antilogarithm, or antilog, of 0.903. An *antilogarithm* is the number that corresponds to a given logarithm. 8, therefore, is the answer to the problem 2×4 . Let us now divide 9 by 3. The logarithm of 9, as we see from the table, is 0.954. The logarithm of 3 is 0.477. Subtracting 0.477 from 0.954, we get 0.477. The table shows that 0.477 is the logarithm of 3. Hence $9 \div 3 = 3$.

Our table of numbers and logarithms to the base ten gives only ten numbers. Mathematicians have prepared tables making it possible to find the logarithm of any number whatsoever. The tables give only the mantissas. We can determine the characteristic in each case by inspection. For example, the logarithm of the number 343 must be between 2 and 3, since $100 = 10^2$ and $1,000 = 10^3$. The logarithm, then, must be 2 and a fraction. The characteristic must be 2. If we look up a five-place table in order to find the mantissa of 343, we observe that it is equal to .53529. Putting together the characteristic 2 and the mantissa .53529, we have the logarithm 2.53529.

In calculations involving arithmetical problems, we can often save a tremendous amount of time by consulting a table of logarithms. Of course we would not use logarithms to get the answer to 4×5 or $72 \div 9$. But suppose we had to perform the various operations in a problem such as:

$$\frac{-2.953 \times 5.913^5 \times \sqrt[3]{5.973}}{49.743 \times 0.35947^3}$$

If the methods of arithmetic were used, this would be a most laborious task. It could be done in a few minutes if we employed logarithms.

Logarithms to the base ten are called *common logarithms*. Other bases have been used. In the so-called *natural logarithms*, the base is 2.71828 . . . , generally indicated by the letter *e*. Natural logarithms serve widely in various types of higher analysis because they lead to comparatively simple formulas.

ALGEBRAIC SEQUENCES AND SERIES

Many events seem to recur in regular sequences. The sun "rises" every day. The planets revolve in their orbits around the sun so regularly that astronomers can calculate their positions years in advance. People have analyzed periodic happenings by means of algebraic sequences and series. A *sequence* is a succession of numbers. A *series* is a sum of numbers in a sequence. The results of these analyses are sometimes used to predict future happenings. We shall briefly consider here the arithmetic and geometric sequences and the binomial series.

The arithmetic series. In an arithmetic sequence, each term, after the first one, is formed by adding a constant quantity to the preceding term. An example of such a sequence is 1, 3, 5, 7, 9, 11, 13, 15, in which 2 is added to each succeeding number of the sequence. If *a* stands for the first number, *d* for the constantly added number, and *n* for the total number of terms, we can represent the arithmetic sequence algebraically by this formula:

$$a, (a + d), (a + 2d) \dots [a + (n - 1)d]$$

Here $[a + (n - 1)d]$ is the *n*th term.

If we add *n* terms together, the sum of the terms is called a series and can be expressed in the formula

$$s \text{ (sum)} = \frac{n}{2} [2a + (n - 1)d]$$

Let us apply this formula to the sum of the terms in the sequence given above: 1, 3, 5, 7, 9, 11, 13, 15. Here there are eight terms in all. The first term is 1. The quan-

tity that is constantly added is 2. Substituting these values for *n*, *a*, and *d*:

$$s = \frac{8}{2} [2 \times 1 + (8 - 1) \times 2]$$

If you work out the arithmetic involved, you will find that the sum of the eight terms is 64. You can verify this by adding the eight terms of the sequence.

The arithmetic series is very useful in various types of calculations. It serves, among other things, in finding the total cost of an item that you are buying on the installment plan. Suppose you buy a piano for \$1,000. You pay \$400 down and agree to pay the other \$600 in 20 monthly installments of \$30 each, plus the interest at 6 per cent on the unpaid balance. Let us apply the arithmetic series to the problem in order to determine the total interest payments that will be required.

The six-per-cent interest means "6 per cent yearly." The installment period is a month, or $\frac{1}{12}$ of a year. The first of these payments is $\frac{1}{12} \times .06 \times \600 (the unpaid balance), or \$3.00. Each month the interest is less than in the preceding month, since the unpaid balance is reduced by \$30. You would pay $\frac{1}{12} \times .06 \times \30 , or \$.15 less interest than the month before. The interest payments, therefore, would be \$3.00 (for the first month), \$2.85 (for the second month), \$2.70 (for the third month) and so on until the 20 installments would be paid. Going back to the formula for the sum of an arithmetic series, we see that *n* (the number of terms) is 20 in this case; that *a* (the first term) is \$3.00; that *d* (the number added to the different terms) is $-.15$. Making the appropriate substitutions in the formula, we have:

$$s = \frac{20}{2} \times [(2 \times 3.00) + (20 - 1) \times -.15]$$

The answer, representing the total interest paid, is \$31.50.

The geometric sequence. In a geometric sequence, each term, after the first one, is formed by multiplying the one before it by the same fixed quantity. A typical geometric sequence is 1, 2, 4, 8, 16, in

which the fixed term used as the multiplier is 2. Algebraically, any geometric sequence can be represented as:

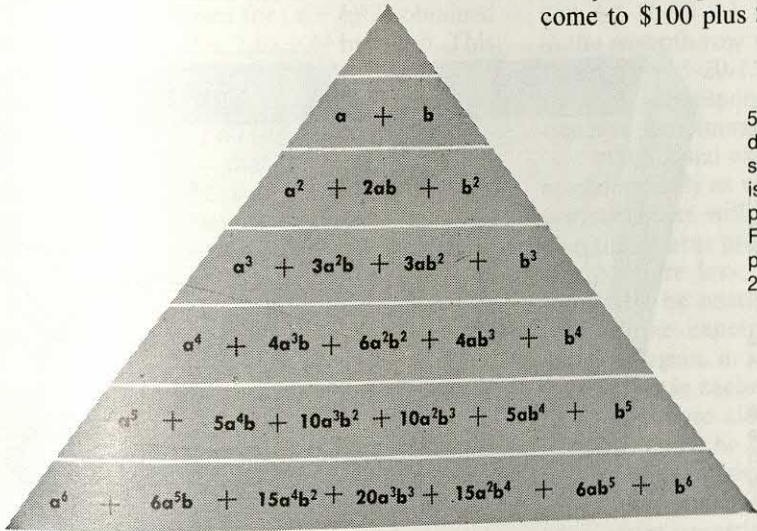
$$a, ar, ar^2 \dots \dots \dots ar^{n-1}$$

where *a* is the first term, *r* the constant multiplier, and *n* the number of terms. The sum of *n* terms of a geometric sequence—a sum called a geometric series—is given by the formula:

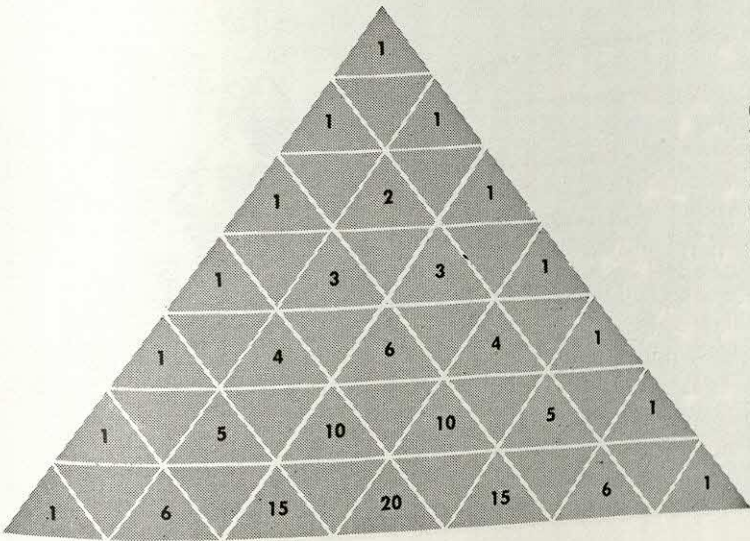
$$s = \frac{ar^n - a}{r - 1}$$

Applying it to the series 1 + 2 + 4 + 8 + 16, you will find that the sum is 31.

The geometric series plays an important part in the mathematics of finance. It is used, among other things, in figuring compound interest. Suppose that we put \$100 in a bank and that the interest is 3 percent, compounded annually. Interest is said to be compounded when it applies, not to the principal alone, but to the principal plus the unpaid interest, periodically added to the principal. At the end of the first year, the principal plus interest would come to \$100 plus \$3.00 (representing the



5. A binomial series, reading downward. Each member of such a series, after the first, is formed by multiplying the preceding member by *a + b*. For example, *a + b*, multiplied by *a + b*, equals *a*² + 2*ab* + *b*².



6. Here the coefficients of the different powers of *a* and *b*, in a binomial series, are arranged in the same order as in Figure 5. A coefficient is a multiplier. For example, 2 is the coefficient in 2*ab*.

interest), or a total of \$103. We could write this as $\$100 \times 1.03$. For the next period—that is, for the second year—we would receive interest of 3 per cent on \$103. Hence the total amount at the end of the second year would be $\$103 + (.03 \times \$103)$. Expressed somewhat differently, the amount in question would be equal to $\$100 \times 1.03^2$.

To repeat, then: at the end of the first year, the total amount we would have in the bank would be $\$100 \times 1.03$; at the end of the second year, we would have $\$100 \times$

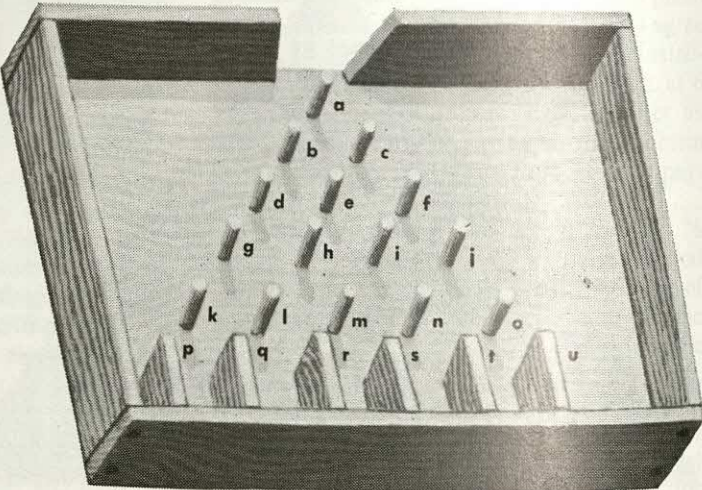
1.03^2 . Following the same procedure, we would have $\$100 \times 1.03^3$ at the end of the third year; $\$100 \times 1.03^4$ at the end of the fourth year and so on. This represents the geometric sequence

$\$100, \$103, \$106.09, \$109.27, \$112.55 \dots$

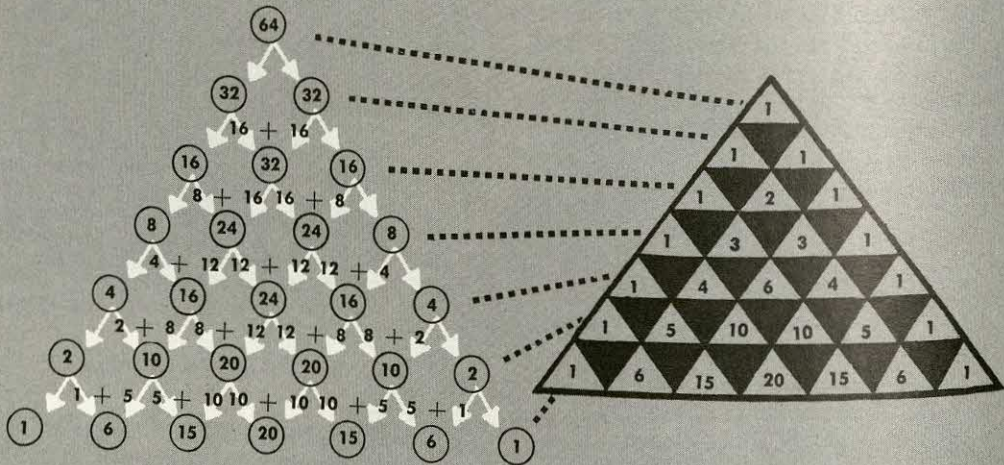
in which 1.03 is the fixed term that is used as the multiplier.

At the end of n periods, the total amount of money in the bank would be $\$100 \times 1.03^n$. For ten interest periods, the

7. The pegs (a, b, c, d, e, f, and so on) in this shallow box are in the exact positions of the numbers in Figure 6. The walls that extend from the bottom pegs form a series of compartments at the bottom of the box.



8. If we drop 64 disks, one by one, in the box shown in Figure 7, they will make their way down the pegs about as indicated. The ratios of the disks hitting the pegs in each row (right) reproduce the triangle in Figure 6.



sum would come to $\$100 \times 1.03^{10}$. Calculating by logarithms, we find that $1.03^{10} = 1.344$. $1.344 \times \$100 = \134.40 ; $\$100$, at 3-per-cent interest, compounded yearly, would be $\$134.40$ at the end of ten years.

The binomial series. A binomial consists of two terms connected by a plus or minus sign. $a + b$, $2x + z$, and $x^2 - y^2$ are all binomials. The binomial series is the sum of terms based on the expansion of the power of a binomial. The series for $(a + b)^1$ is $a + b$. For $(a + b)^2$, we obtain the series by multiplying $a + b$ by $a + b$, giving $a^2 + 2ab + b^2$. The series for $(a + b)^3$ is obtained by multiplying $a^2 + 2ab + b^2$ by $a + b$. This gives $a^3 + 3a^2b + 3ab^2 + b^3$. We can arrange these binomial series and also those for $(a + b)^4$, $(a + b)^5$, and $(a + b)^6$ as in Figure 5.

Suppose now that, instead of the different powers of a and b , we give only the coefficients—numerals or letters used as multipliers—and reproduce the triangle in this modified form as shown in Figure 6. You will note that we add the number 1 at the peak of the triangle.

Let us now put a series of pegs in a shallow box, in the exact positions of the numbers of the numerical triangle, as shown in Figure 7. The pegs are to be just far enough apart so that a small disk will be able to pass between them. We cut out a small section of one side of the box (which will be the top), as indicated in the figure, and we also build a series of walls extending from each of the bottom pegs, so as to form a series of compartments.

We keep the box in a tilted position so that when a disk is dropped through the gap at the top, it will make its way through the maze of pegs to one of the bottom compartments. Now we drop 64 disks one by one into the box through the gap. We can expect that half of the disks that hit a peg will fall to the left of it and the other half to the right. Hence 32 disks should drop to the left of peg a and 32 to the right of the peg. Of the 32 that fall to the left and strike peg b , 16 should fall to the left of b and should strike peg d ; 16 should fall to the right of peg b and should hit peg e . Of the 32 disks that hit peg c , 16 should strike peg e and 16 should strike peg f . That

means that 32 disks in all will hit peg e . We can indicate how the disks should fall on their way to the bottom compartments by the diagram in Figure 8.

As we have seen, 32 disks should strike peg b in the second row and 32 should strike peg c . 32 and 32 are in the ratio 1-1. In the third row, 16 should strike d ; 32 should strike e and 16 should strike f . The ratio between 16, 32 and 16 is 1-2-1. Going down the rows, the ratios of the disks striking the different pegs would be 1-3-3-1 in the fourth row, 1-4-6-4-1 in the fifth row and 1-5-10-10-5-1 in the sixth. The disks in the seventh row would follow the distribution 1-6-15-20-15-6-1. Note that these ratios all correspond exactly to the coefficients of the binomial series in Figure 6.

In an actual experiment, the disks will not fall exactly as we have indicated. Some compartments will have one or two more than the number predicted. Others will have one or more less. Yet in every case the result will be nearly that which was forecast. If the experiment is repeated over and over again, in a great number of trials, the number in each compartment will agree more and more closely with the expected number. Thus the coefficients of the binomial series provide an effective means for calculating probabilities when the chances of an event occurring are even.

USED IN STATISTICS AND PROBABILITY

To determine probabilities when the chances are not even, other algebraic analyses have been made. The subject that deals with such analyses is called *statistics*. It is used in many fields, including insurance. Insurance companies can, for example, predict with fair accuracy life expectancies and how much in premiums and interest they will collect during the lifetime of a person with insurance. To make insurance work, the company must collect enough to be able to pay the insurance when the client dies. Probability and compound interest, as developed by algebra, are therefore the bases on which insurance is built. Thus algebra, which began by examining the relations of arithmetic operations, has become an interpreter of our experience and a guide for the future.



Thomas R. Taylor, Photo Researchers



John R. Ross, Photo Researchers

Examples of geometry can be found in infinite variety throughout nature, science, and art. Above left: the shell of a chambered nautilus forms a spiral whose dimensions increase uniformly. Above right: the Great Pyramid at Giza is a testimony to ancient peoples' knowledge of geometry.

PLANE GEOMETRY

by Howard F. Fehr

When we pass from arithmetic and algebra to geometry, we enter a world of shapes occurring in space—a world of points, lines, surfaces, and solids. We study the properties of these shapes and the relations between them. We learn to measure them. At the outset, geometry was used to solve specific problems, but in the course of its development it became a thoroughly abstruse subject. However, this abstruse branch of mathematics can often be put to practical use, as we shall see.

The beginnings of geometry go back far into prehistory. As the population of a given region grew, the natural dwelling places available did not suffice. It became necessary to build shelters, big enough to house families and strong enough to withstand winds, rain, and storms. To make a shelter the proper size, a person had to compare lengths. Thus the roof had to be higher above the ground than the top of the head of the tallest person.

The ancient Babylonians were pioneers in this branch of mathematics. The land between the Tigris and Euphrates rivers, where the Babylonians dwelt, was originally marshland. Canals were built to drain the marshes, and to catch the overflow of the rivers. For the purposes of canal construction, it was necessary to survey the land. In so doing, the Babylonians developed rules for finding areas. These rules were not exact, but the results they gave sufficed for canal construction.

In Egypt, the people who had farms along the banks of the Nile River were taxed according to their holdings. In the rainy season, the river would overflow its banks and spread over the land, washing away all landmarks. It became necessary, therefore, to remeasure the land so that each owner would have his rightful share. After the floods had subsided, specially trained men, called rope-stretchers, would establish new landmarks. They would use ropes knotted at equal intervals so that they could measure out desired lengths and divide the land into triangles, rectangles, and trapezoids.

They devised practical rules for the areas of these figures. The rules were of the rough-and-ready variety and were often inexact. We know today, for example, that

the area of any triangle is one-half the product of its altitude, or height, and its base. The Egyptians erroneously gave this area as one-half the product of the base and a side. However, most of the triangles used in their surveying work were long and narrow (Figure 1); and in such triangles there is not too much difference in length between the side and the altitude. Hence the results of the Egyptians' calculations served as a pretty fair basis for the allotting of land and the taxation of landowners.

GEOMETRY BECOMES A DISCIPLINE

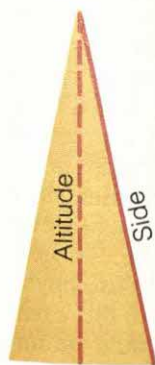
The Greeks called the early Egyptian surveyors *geometers*, or earth-measurers (from the Greek *ge*: "earth" and *metria*: "measurement"). The geometers found out many facts about triangles, squares, rectangles, and even circles. These facts became a body of knowledge that the Greeks called *geometry*, or "the study of the measurement of the earth." Geometry today involves much more than it did at that early stage; yet it is still concerned with the sizes, shapes, and positions of things.

The Greeks made important advances in the field of geometry. They not only corrected many of the faulty rules of the Egyptians, but also studied the different geometrical figures in order to work out relationships. Thales, a Greek mathematician who lived 2,500 years ago, discovered that no matter what diameter one draws in a circle, it always bisects the circle—that is, cuts it into two halves (Figure 2). He also noticed that if two straight lines cross each other, the opposite angles are always equal, no matter at what angle the lines cross (see

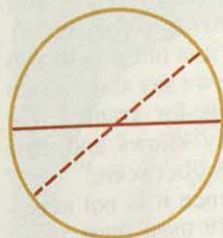
a and *b* in Figure 3). This was the beginning of the study of figures for the sake of discovering their properties rather than for practical use. The Greeks changed geometry from the study of land measurement to the study of the relations between different parts of the figures existing in space. This is what geometry means today.

After Thales, other Greek mathematicians discovered and proved facts about geometric figures. They set forth these facts in statements called *theorems*. They also devised various instruments for drawing figures. By custom, the only instruments allowed in the formal study of geometry were an unmarked straightedge, or ruler, for drawing straight lines and a compass for drawing circles and transferring measurements (Figure 4).

The Greeks proposed various construction problems, to be solved with only the straightedge and the compass. Among these problems were the following: (1) Construct a square whose area exactly equals that of a given circle—called squaring the circle. (2) Construct the edge of a cube whose volume will be exactly twice the volume of a given cube—called duplicating the cube. (3) Construct an angle equal to exactly one-third of a given angle—called trisecting the angle. For over twenty-two centuries, mathematicians at-

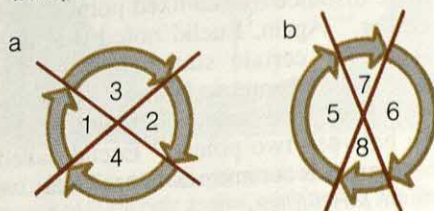


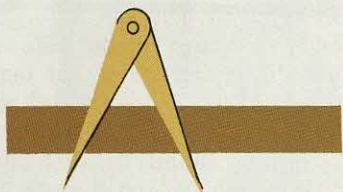
1. In this long and narrow triangle, each side is not a great deal longer than the altitude. The Egyptians used such triangles in surveying.



2. Each of the two diameters shown in the drawing cuts the circle in halves.

3. If two lines cross, the opposite angles are always equal. In *a*, angle 1 = angle 2 and angle 3 = angle 4. In *b*, 5 = 6 and 7 = 8.





4. By custom, the ancient Greeks used only an unmarked straightedge and compass in geometry.

tempted to solve these problems, without success. Finally, in the nineteenth century, it was proved that it is impossible to square the circle, duplicate the cube, or trisect an angle if one uses only the straightedge and compass. It is possible, however, to make these three constructions with specially-designed instruments. Such constructions fall in the domain of higher geometry.

EUCLID—ORGANIZER OF GEOMETRY

By the fourth century B.C., there had grown up a vast body of facts concerning geometric figures, but for the most part these facts were unrelated. There were many theorems about triangles and circles, some about similar figures and areas, but no orderly arrangement. The learned Greek mathematician Euclid, who taught at the Museum of Alexandria, in Egypt, about 300 B.C., was the first man to apply a logical development to the mathematical knowledge of his time. He presented this development in his *Elements of Geometry*.

Euclid realized that it is not possible to prove every single thing we say and that we must take certain things for granted. He assumed that everybody knows and uses properly such words as "between," "on," "point," and "line"; hence it is not necessary to define them. He used these undefined terms to give definitions of various figures. Thus he defined a circle as "a closed curved line every point of which is the same distance from a fixed point called the center." Again, Euclid noted that one cannot prove certain statements of relations between geometric figures. An example would be: "Only one straight line can be drawn between two points." Euclid called such statements common notions. Today we call them *postulates*.



5. If two sides of a triangle are equal, the angles opposite these sides are equal; that is, if $I = II$, $1 = 2$.

Euclid used undefined terms, definitions, and postulates to prove *theorems* about geometric figures. A theorem is a statement that gives certain facts about a figure and that concludes from these facts that a certain other fact must be true. A typical theorem is "If two sides of a triangle are equal, the angles opposite these sides must be equal" (Figure 5). The theorem states the facts that (1) there is a triangle and that (2) two sides of the triangle are equal. It then draws the conclusion that two of the angles of the triangles are equal. Once a theorem is proved, it can be used to prove other theorems.

Euclid built up a logical chain of theorems that introduced order in what had been a chaos of more or less unrelated facts. Besides organizing a vast body of knowledge about geometric figures, he introduced a method of treatment that became a model for the development of other branches of mathematics and pure science. This method is as valid today as ever.

TWO DIMENSIONS

The first branch of geometry we shall consider is plane geometry—the study of points, lines, and figures occurring in planes. Just what do we mean by these terms?

A *point* is the simplest element in geometry. It has neither length nor width nor thickness, which is another way of saying that it has no dimensions at all. We can represent a point by a dot, made with a lead pencil or a piece of chalk. Such a dot is not a geometric point but a physical point, since it has length, width, and thickness, however small these dimensions may be. In geometric constructions, we have to use physical points, such as pencil dots, to represent geometric points, because it would be im-

possible for us to set down on paper a point with no dimensions.

If there are two different points, the shortest distance between them is a *straight line*. This line segment has only one dimension, called *length*. It does not have width and thickness. A straight line that we draw on paper with a pencil does have width and thickness. Hence, when we draw a line in constructing a geometric figure, we are again giving a physical representation of a geometric element.

If we were confined to a world having only one dimension, such as length, we would have a rather dull time of it. We would be points on a line, being able to move only forward and backward and always bumping into points ahead of us or behind us. In Figure 6, the points *A* and *B* and the segment *CD* are all parts of the line shown in the figure and must always stay within the line.

Suppose now that we selected a point *P* outside the line. The lines drawn through the point *P* and meeting the original line create a series of figures existing in a plane. A *plane* is a surface having the two dimensions of length and width (Figure 7). The surface of a table top is a plane. A continuation of the surface would represent part of the same plane. If we were points in a two-dimensional world, we could move freely in any direction, except out of the plane. Our world would have other points like ourselves, and also lines. There would also be a great variety of figures—made up of combinations of points and lines—figures such as triangles, squares, circles, and so on.

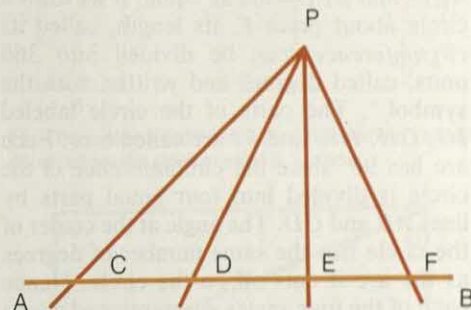
ANGLES IN PLANE GEOMETRY

The name *ray* is given to the part of a line that starts at a given point. A plane figure formed by two rays having the same starting point is called an *angle*. In Figure 8, *AB* and *BC* are two rays with the same starting point, *B*. The angle formed by the two rays is *ABC*. You will note that letter *B*, standing for the starting point, is inserted between letter *A*, on one of the two rays and letter *C*, on the other ray. That is how angles are always indicated.

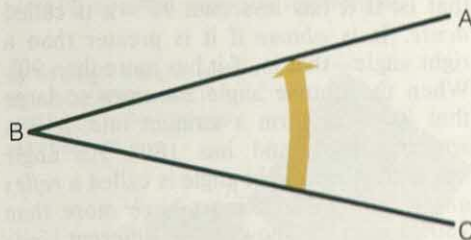
If two lines meet so that all the angles



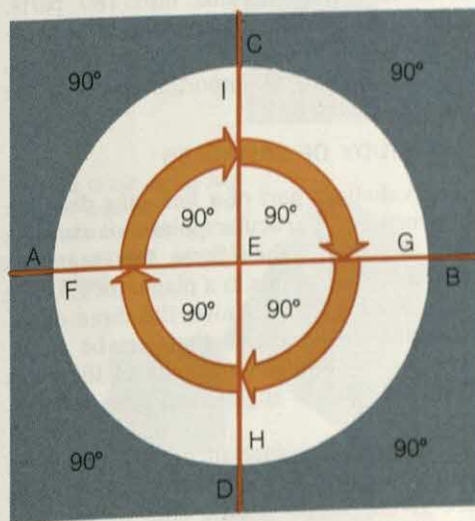
6. Points *A* and *B* and the length *CD* are parts of the line above and must always stay within the line.



7. Lines drawn from *P* to the line *AB* create a number of figures (such as *PCD*) occurring in a plane.



8. In the above, *AB* and *BC* are two rays having the same starting point, *B*. They form angle *ABC*.



9. Lines *AB* and *CD* are perpendicular to each other; the angles at *E* are right angles, each having 90° . The arcs cut off on the circle also each have 90° .

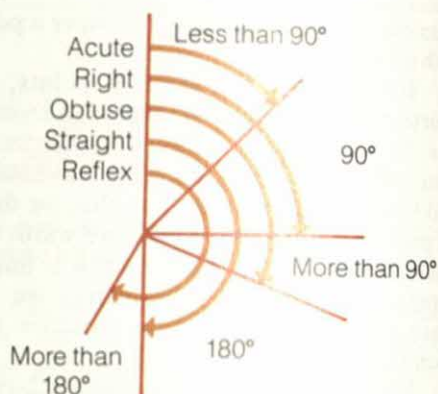
formed are equal, the lines are said to be *perpendicular* and the angles are called *right angles*. In Figure 9, AB is perpendicular to CD and the four angles— AEC , BEC , AED , and BED —are all equal. If we draw a circle about point E , its length, called its *circumference*, can be divided into 360 units, called degrees and written with the symbol $^\circ$. The parts of the circle labeled IG , GH , HF , and FI are called *arcs*. Each arc has 90° since the circumference of the circle is divided into four equal parts by lines AB and CD . The angle at the center of the circle has the same number of degrees as the arc it cuts off on the circle. Hence each of the four angles we mentioned above has 90° . In other words, a right angle has a measurement of 90° .

If an angle is less than a right angle—that is, if it has less than 90° —it is called *acute*. It is *obtuse* if it is greater than a right angle—that is, if it has more than 90° . When the obtuse angle becomes so large that its sides form a straight line, it is a *straight angle* and has 180° . An angle larger than a straight angle is called a *reflex angle*: of course it must have more than 180° . Figure 10 shows these different kinds of angles. Angles can be measured by the instrument called the *protractor*. It consists of a semicircle divided into 180 parts, each part representing one degree of angle at the center (Figure 11). As we shall see, angles play an all-important part in the study of geometry.

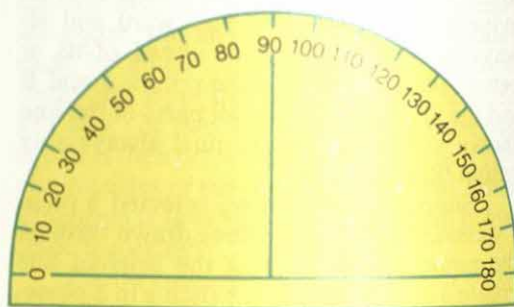
THE STUDY OF TRIANGLES

A distinct part of a line—the distance between two particular points—is called a *line segment*. When three line segments connect three points in a plane, they form a *triangle*. Figure 12 shows the three different kinds of triangles that can be made. There are literally thousands of theorems about the sides, angles, and lines in triangles.

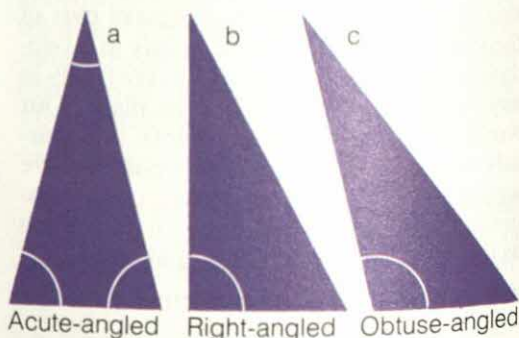
One of the first theorems proved in plane geometry is “If three definite lengths are given, such that the sum of any two lengths is greater than the third length, it is possible to use the lengths in making a triangle that will have a definite size and



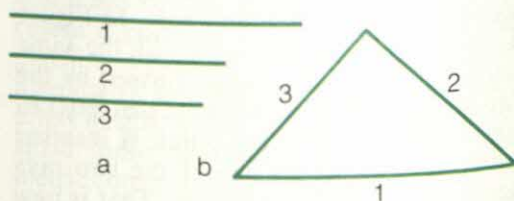
10. The different types of angles, from acute to reflex.



11. Angles are measured by a device called a protractor.



12. Above are the three kinds of triangles.



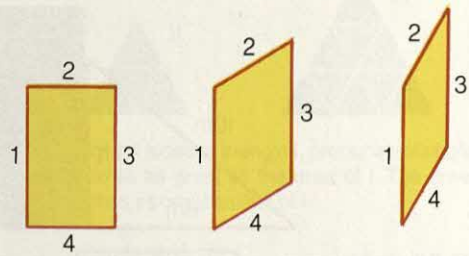
13. In a three line segments (1, 2, and 3) are shown. In b, these segments are joined to form a triangle.

shape" (Figure 13). Since the shape never varies, a construction built in the form of a triangle will be rigid and will not "give." Because of this property, interlocking triangles are used in all bridge and building design in order to prevent a structure from collapsing. A figure of four sides, each of a definite length, could have many different shapes, as shown in Figure 14. A construction built in this shape would be collapsible. Hence it cannot be used in rigid construction unless it is braced by a diagonal. Each diagonal makes two triangles of a four-sided figure (Figure 15), and each of these triangles is rigid.

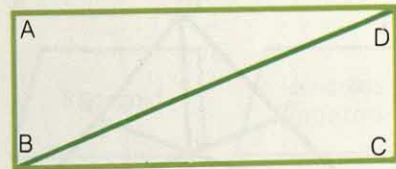
The most famous and perhaps the most important theorem in plane geometry is one dealing with a right triangle. A right triangle is a triangle having a right angle. It is called the Pythagorean theorem, after its discoverer, the Greek philosopher Pythagoras, who lived in the sixth century B.C. This theorem states that "The sum of the squares on two sides of a right triangle is equal to the square on the hypotenuse (the side opposite the right angle)." In the right triangle in Figure 16, the sides are 3, 4, and 5 units in length. The side with 5 units is the hypotenuse. We draw the three large squares as shown: the first with a side (AB) consisting of 3 units, the second with a side (BC) consisting of 4 units, the third with a side (AC) consisting of 5 units. According to the Pythagorean theorem, $(AB)^2 + (BC)^2 = (AC)^2$. In this case $3^2 + 4^2 = 5^2$, or $9 + 16 = 25$. We can verify this by counting the small squares, each with a side a unit long, in the three large squares. There are 9 small squares in square $ABED$, 16 in square $BCGF$, and 25 in square $ACIH$.

Figure 17 shows how the two smaller squares on the sides of a right triangle can be cut up so as to form the square on the hypotenuse. This is another confirmation of the Pythagorean theorem.

It follows from this theorem that if a triangle has sides such that the sum of the squares of the two smaller sides is equal to the square of the largest side, the angle opposite the largest side is a right angle. This theorem has various practical applications. The carpenter uses it to see whether

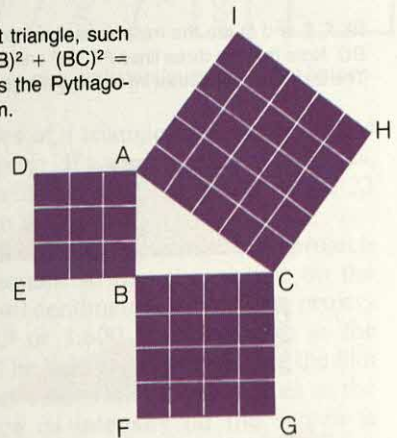


14. In these four-sided figures, the sides marked 1 are all equal; so are the sides marked 2, 3, and 4.

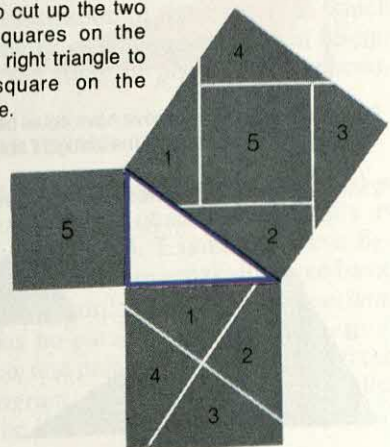


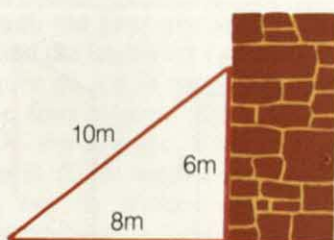
15. To brace the four-sided construction, the diagonal piece BD is inserted, making two rigid triangles.

16. In a right triangle, such as ABC , $(AB)^2 + (BC)^2 = (AC)^2$. This is the Pythagorean theorem.

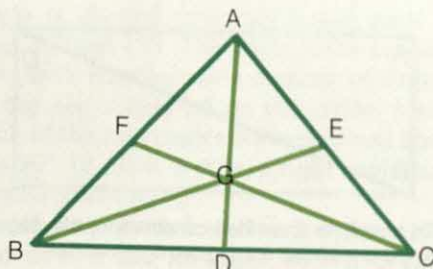


17. How to cut up the two smaller squares on the sides of a right triangle to form a square on the hypotenuse.

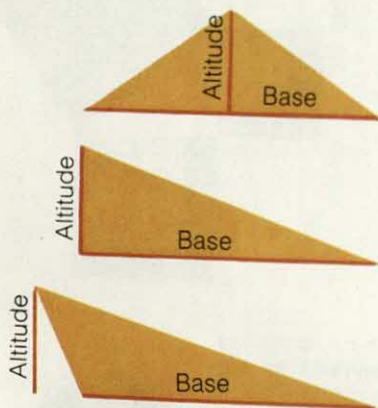




18. If the wall is perpendicular to the floor, the triangle will fit snugly.

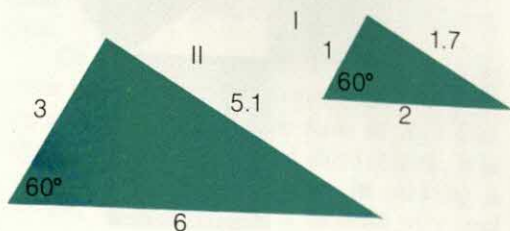


19. F , E and D are the midpoints of sides AB , AC , and BC . Note that the three lines FC , BE and AD meet at G . The text points out other interesting things about G .



20. The three triangles above have equal bases and altitudes. Hence the areas of the triangles are also equal.

21. Triangles with equal corresponding angles are similar.



a wall is perpendicular to the floor. If boards of 6, 8, and 10 meters for example, are joined together as shown in Figure 18, the angle between the 6- and 8-meter lengths must be a right angle, since $6^2 + 8^2 = 10^2$, or $36 + 64 = 100$. If the wall is truly perpendicular to the floor, the triangle will fit snugly.

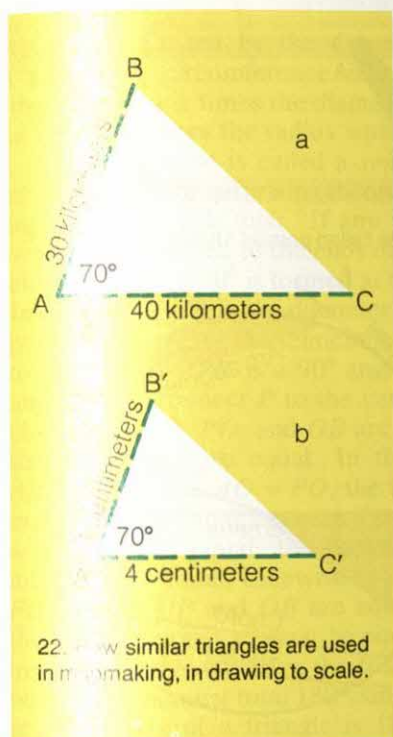
The Pythagorean theorem is one of many that reveal an unexpected and important relationship. Here is another instance of such a theorem. In the triangle ABC , in Figure 19, D , E , and F are the midpoints of the sides BC , AC , and AB . The points A , B , and C are the vertexes of the triangle. If we connect the vertex A to the midpoint, D , of the opposite side, BC , the line AD is called a *median*. Let us draw the two other medians of the triangle, CF and BE . We learn what we had not suspected—that these three medians all pass through the same point G , inside the triangle. We also learn that G on any of the medians is two-thirds the distance from the vertex to the opposite side. Plane geometry offers proofs of these statements. Here is another interesting fact about G . If we cut out a triangle of cardboard and draw the three medians, as in Figure 19, we can balance the triangle on the blunt end of a lead pencil if we put this end directly under G . This point is called the *center of gravity*.

EQUAL AND SIMILAR TRIANGLES

When triangles have the same size, or area, they are called *equal triangles*. All triangles with equal bases and altitudes are equal although they may have many different shapes (Figure 20).

Some triangles have the same shape, but are different in size. They are known as *similar triangles*. The corresponding angles of similar triangles are equal and their corresponding sides are always in the same ratio. Thus the two triangles I and II shown in Figure 21 are similar because the corresponding angles are equal. Each side of triangle II is 3 times as great as the corresponding side of triangle I.

Similar triangles are used in drawing to scale. In making a map, for example, we represent a large area of land on a small

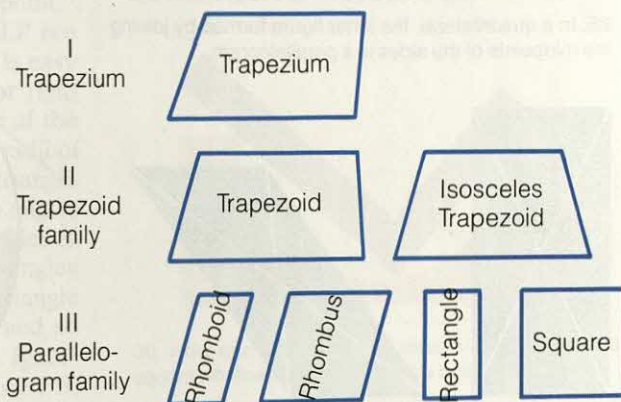


piece of paper. Suppose site *A*, in Figure 22*a*, is 30 kilometers from site *B* and 40 kilometers from site *C* and that the angle *CAB* is 70° . We are to show these sites on a map, where the scale is to be 1 centimeter = 10,000 meters. First, using a protractor we draw an angle of 70° . Thirty kilometers is equal to 30,000 meters, since there are 1,000 meters in a kilometer. Since one centimeter equals 10,000 meters, to represent 30 kilometers, or 30,000 meters, we need three centimeters. To represent 40 kilometers, we need four centimeters. We measure off three centimeters on one side of the angle and four centimeters on the other side of the 70° angle. Joining the ends of these segments, we have a map (Figure 22*b*) representing the triangular area formed by the three sites.

All maps, models, and photographs are similar to the original objects they represent. Hence the angles in these representations are exactly the same size as in the originals, and all lines are changed in the same ratio. Their areas will also have a definite ratio. They will vary as the squares of the corresponding sides. If we double each



23. If we double the length of the sides of triangle I, producing triangle II, the area of II will be four times as great as the area of I. The area of triangle III, above, is nine times as great as that of I.



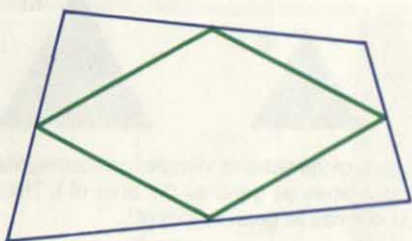
24. The three basic types of quadrilaterals, or four-sided figures.

of the sides of a triangle, the area will be 4 times as great. If we triple each of the sides, the area will be 9 times as great. Figure 23 shows that this is so.

If film one centimeter square is projected on a screen so that the picture on the screen is 40 centimeters square, the projection is 40^2 or 1,600 times as large as the original. The light used in projecting the film must cover 1,600 times as much area as the film; hence its intensity on the screen is only $1/1600$, as great as at the film. This is a striking illustration of the manner in which the geometry of similar figures can be applied to the study of photographic phenomena.

QUADRILATERALS

A figure with four sides is called a *quadrilateral*. The quadrilateral family is shown in Figure 24. Examining these figures, we see that there are really three basic types of quadrilaterals: the *trapezium*, which has no parallel sides; the *trapezoid*, which has one pair of parallel sides; and the *parallelogram*, with two pairs of parallel sides. The *isosceles trapezoid* is a trapezoid



25. In a quadrilateral, the inner figure formed by joining the midpoints of the sides is a parallelogram.

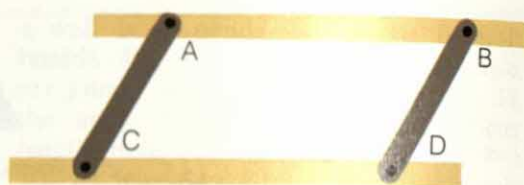


26. The corresponding sides of these parallelograms are equal. The diagonals of both bisect each other.

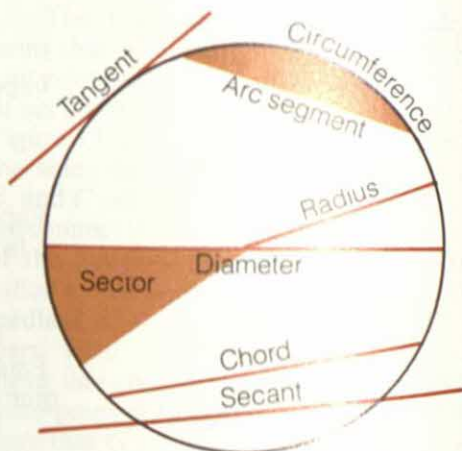
whose two nonparallel sides are equal. A *rhomboid* is a parallelogram with no right angles. A *rhombus* is a rhomboid with four equal sides. A *rectangle* is a parallelogram whose angles are all right angles. A *square* is a rectangle with equal sides. In all these figures, if we connect the midpoints of the sides, as in Figure 25, the inner quadrilateral that is formed will always be a parallelogram.

In a parallelogram, the diagonals bisect each other, no matter how we distort the figure (Figure 26). This is a good example of an *invariant*—a property of a figure that remains true under all distortions. Another invariant is “The opposite sides of a parallelogram are equal.”

The draftsman makes use of this invariant in the instrument called the parallel rulers (Figure 27). It consists of two straightedges, which are joined by two rods AC , and BD in such a way that $AC = BD$. This device is flexible; hence AB can be at varying distances from CD . However, no matter how AB is moved it always remains parallel to CD . Hence, using this device,



27. The device called parallel rulers.



28. The principal parts of a circle are shown above.

we can draw a parallel to a given line at any accessible point in a plane.

The study of triangles and quadrilaterals forms much of the subject matter of geometry. All other polygons—closed figures having three or more angles and therefore sides—can be divided into triangles and quadrilaterals by drawing diagonals from the vertices of the polygon. This then means that the basic rules for determining the perimeter and area of triangles and quadrilaterals can be applied to the study of other polygons.

THE CIRCLE

The study of the circle is also important. In Figure 28, we show its important parts. You will note that the closer a *chord* gets to the center of the circle, the larger it becomes. The *diameter* is really a chord that passes through the center. It consists of two radii joined together so as to form a straight line. To measure the length of a circle—its *circumference*—we find the number of diameters in it. This number is about $3\frac{1}{7}$ or, more accurately, 3.1416; its exact

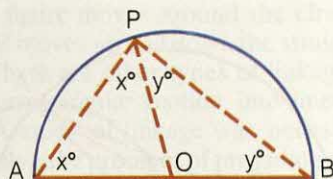
value is indicated by the Greek letter π ("pi"). The circumference of the circle, then, is πd or π times the diameter. Its area is πr^2 , or π times the radius squared.

A half circle is called a *semicircle*. A simple and quite surprising theorem involving a semicircle is this: "If any point on a semicircle is joined to the ends of the diameter, an angle of 90° is formed at the point." In Figure 29, AB is the diameter and P is a point anywhere on the semicircle. It is easy to prove that APB is a 90° angle, or right angle. First connect P to the center of the circle (O). AO , PO , and OB are all radii of the circle and are equal. In the triangle APO , since side $AO = PO$, the two angles marked x° are equal, because if two sides of a triangle are equal, the opposite angles must also be equal. Likewise in the triangle POB , sides OP and OB are equal, and so the angles marked y° must be equal. Ignoring the line OP , we have the triangle APB , whose angles must total 180° , since the sum of the angles of a triangle is 180° . There are two x° angles and two y° angles in the triangle APB ; hence one x° angle and one y° angle must give half of 180° , or 90° . Since the angle APB is composed of an x° angle and a y° angle, it must be equal to 90° ; it must be a right angle.

There are various applications of this theorem. For example, a pattern maker can determine if the core box shown in Figure 30 is a true semicircle. He places a device called a carpenter's square in the box. If it makes firm contact at three points, as shown, he knows that the pattern will give a true semicircle.

An angle at the center of a circle has as many degrees as the arc that it intercepts, or cuts off, on the circle. Or, as a mathematician would say, "A central angle is measured by its intercepted arc." In Figure 31, the obtuse angle at O is equal to 110° , and the arc AB it intercepts is also equal to 110° .

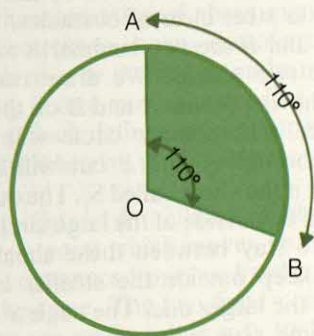
An angle whose vertex is on the circumference of a circle and which cuts off or intercepts an arc is called an *inscribed angle* (angle ABC in Figure 32). No matter where we place the vertex B in this arc, the angle will remain the same size. In other words,



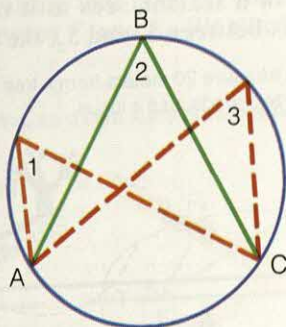
29. If point P on the semicircle is joined to the diameter at A and B , angle APB will be a 90° angle.



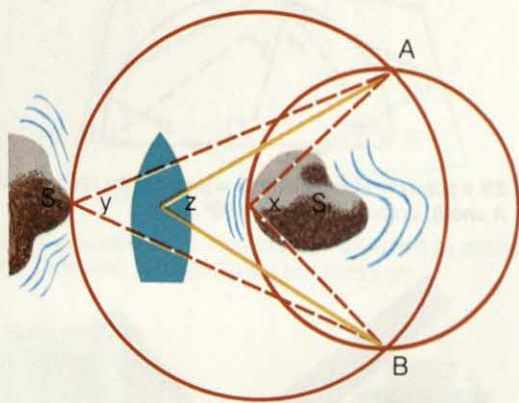
30. How one determines by means of a carpenter's square whether a core box gives a true semicircle.



31. Both the angle at O and the arc that it intercepts (AB) on the above circle are equal to 110° .



32. The angles marked 1, 2 and 3 intercept the same arc (AC) in this circle; therefore they are equal.



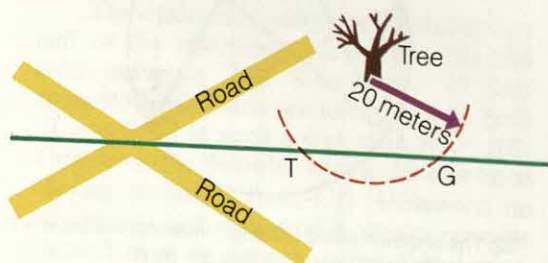
33. The ship can be steered so as to avoid the shoals S_1 and S_2 .

all inscribed angles intercepting the same arc are equal.

This invariant is used by navigators in order to steer clear of obstacles. In Figure 33, A and B are two landmarks and S_1 and S_2 are two shoals. We draw two circles, each having points A and B on the circumference. The smaller circle will not only pass through A and B but will also have within it the shoal called S_1 . The other shoal, S_2 , will be outside of the large circle.

To stay between these shoals, a ship must keep outside the smaller circle and inside the larger one. The angle x is always the same size, wherever it may be in the smaller circle, since it always intercepts the same arc; the angle y is always the same size for the same reason. The navigator, by the use of a sextant, sees to it that as the ship sails between S_1 and S_2 , the angle z is

34. To find a treasure 20 meters from a tree and equally distant from two roads, use a locus.



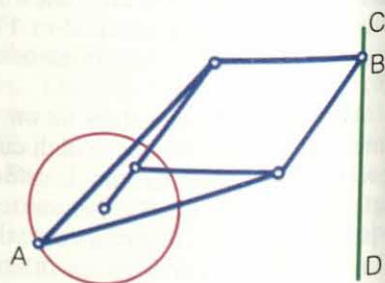
smaller than the angle x (which keeps him outside the smaller circle) but greater than the angle y (which keeps him inside the larger circle). Thus the ship avoids both shoals.

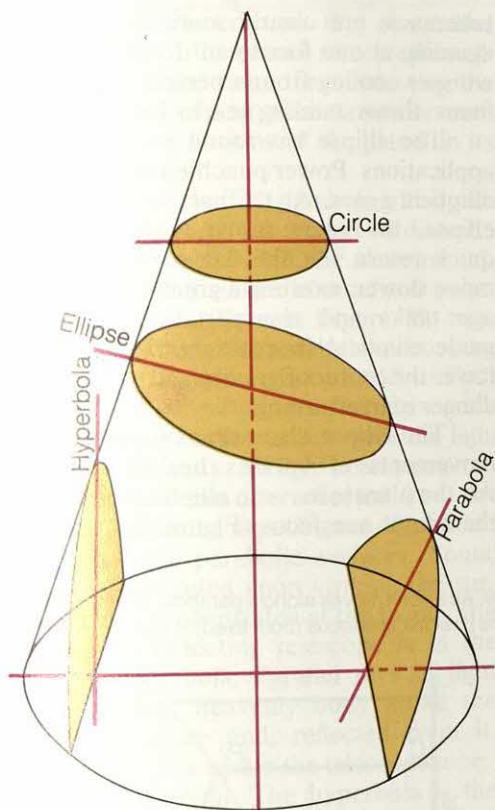
THE PATH OF A MOVING POINT

It is often necessary in plane geometry to determine the path that a point describes in a plane when it moves according to a fixed rule. If, for example, a point must always remain 3 centimeters from a fixed point, it travels in a circle around the fixed point. The mathematician gives the name *locus* to the path described by a point. (*Locus* means "position" in Latin.) By studying the paths of moving points, we determine how machine parts move and how heavenly bodies appear to move.

We can illustrate the use of the locus by a very simple treasure-hunt problem. A treasure is reported to be buried 20 meters from an oak tree and also equally distant from two intersecting roads (Figure 34). Since it is 20 meters from the oak tree, it is somewhere on the circumference of a circle with a radius of 20 meters and with the tree as the center. Now we also know that the treasure is equidistant between the two intersecting roads. We may think of it as a moving point remaining at the same distance from each road. It can be proved in geometry that a point moving so as to be equally distant from the sides of an angle, traces a line that bisects the angle. Hence the treasure must be on a line bisecting the angle made by the two roads. This line cuts the circle, as shown in Figure 34, at two places, T and G . The treasure, therefore,

35. Peaucellier's Cell changes circular motion to straight-line motion. It is a type of linkage.





36. The conic sections—circle, ellipse, parabola and hyperbola—produced as planes cut a cone.

must be at one or the other of these two points.

Machines that are designed to trace moving points are called *linkages* because they consist of linked bars. A pair of compasses is the simplest linkage. The path it traces is a circle. Another linkage, called Peaucellier's Cell, changes circular motion into straight-line motion (Figure 35). As A

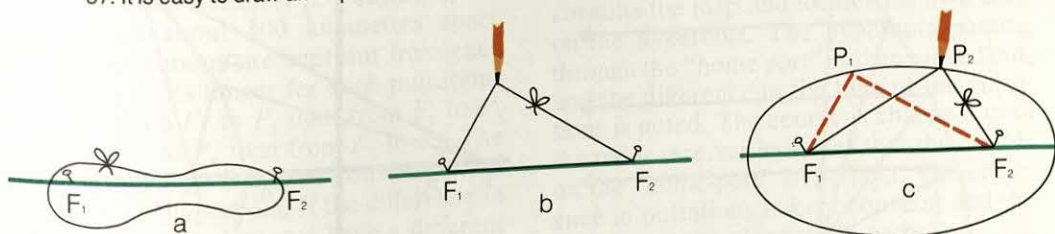
in the figure moves around the circle, the point B moves up and down the straight line CD . There are other types of linkages that transform circular motion into linear motion. A study of linkage was necessary to help solve the problem of providing smooth motion in a locomotive, where the straight-line motion of the drive shaft had to be converted into the circular motion of the wheels.

CONIC SECTIONS

The circle is the most common type of curve, but there are other kinds. The Greek geometers noticed very early that when a cone was cut by planes at different angles, the intersections gave different kinds of curves: circles, ellipses, parabolas, and hyperbolas (Figure 36). Because the three latter kinds of curves were first described in connection with cones, they were called *conic sections*. Apollonius of Perga, who lived in the third century B.C., wrote a treatise on the properties of these curves. In more recent times, it was discovered that they could also be defined as paths made in a plane by points moving according to certain rules. Such definitions are particularly meaningful when we put the curves to practical use.

The ellipse. An ellipse is the path traced by a point which moves so that the sum of its distances from two fixed points is always the same. The two fixed points are called *foci*, or *foci*. It is easy to draw an ellipse, using the method illustrated in Figure 37. First we insert thumbtacks at two fixed points, F_1 and F_2 . We then take a piece of string that is larger than the distance between F_1 and F_2 . We attach one end

37. It is easy to draw an ellipse, using two thumbtacks and string, as shown in the drawings below.



of the string to the thumbtack at F_1 and the other to the thumbtack at F_2 (Figure 37a). We draw the string taut and insert a pencil as shown in Figure 37b. As we move the pencil, its point will trace an ellipse (Figure 37c). The sum of the distances from the moving pencil point to the fixed points will remain constant. In figure 37c, $P_1F_1 + P_1F_2 = P_2F_1 + P_2F_2$.

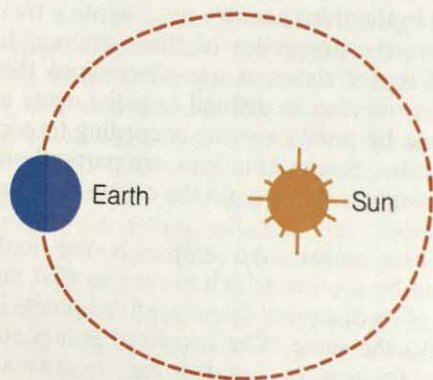
If a billiard table were elliptical in shape, any ball hit from one focus would rebound through the other focus. In an elliptical room, any sound issuing from one focus will be reflected by the walls to the other focus. This is the principle of the "whispering gallery." The elliptically shaped Mormon tabernacle in Salt Lake City, Utah, is an example. The foci in the

tabernacle are clearly marked. A person standing at one focus can distinctly hear a whisper coming from a person at the other focus; those standing nearby hear nothing.

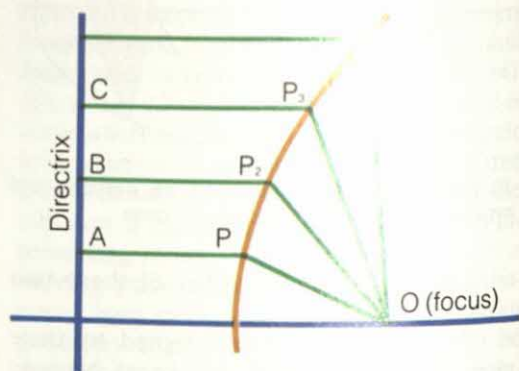
The ellipse has found many practical applications. Power punching machines use elliptical gears. At the narrow ends of the ellipse, the gears move faster, giving a quick return. At the flat parts, the gears move slower, exerting a greater force. Storage tanks and transportation tanks are made elliptical in cross section so as to lower the center of gravity and to lessen the danger of overturning.

The ellipse also serves to explain the movements of various heavenly bodies. All the planets move in elliptical orbits with the sun at one focus (Figure 38). A planet

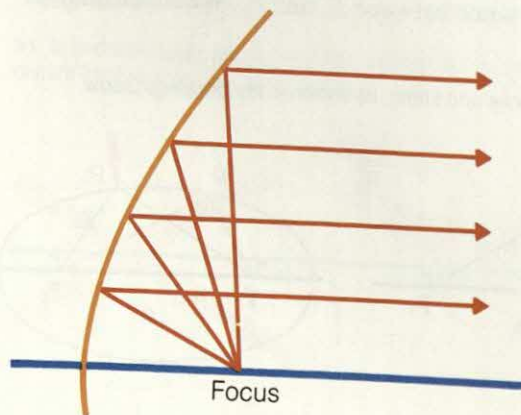
38. As the planet earth travels around the sun, its orbit is an ellipse, with the sun at one focus.



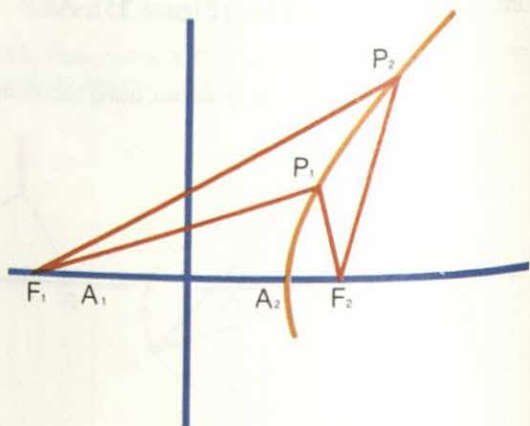
39. As a point moves along a parabola, its distance from the directrix and focus (both fixed) is equal.



40. The light beams emanating from the focus of the searchlight that we show below are reflected from the parabolic surface in a series of parallel rays.



41. P_1 and P_2 are two positions of a point moving along a hyperbola. The distance to F_1 minus the distance to F_2 always equals the distance A_1A_2 .



moves along the orbit so that the radius from the sun to the planet sweeps through equal areas in the same time. Knowing the elliptical orbit of any planet, astronomers can predict the position of the planet in its orbit at any time.

The parabola. The parabola is the path of a point which moves so that its distance from a fixed line, called the *directrix*, always equals its distance from a fixed point, called the focus. Thus in Figure 39, as a point moves along the parabola, occupying positions P_1 , P_2 and P_3 in turn, $AP_1 = P_1O$; $BP_2 = P_2O$; $CP_3 = P_3O$. A reflecting searchlight has a parabolic surface with the light source at the focus. All light beams emanating from the focus are reflected from the parabola in parallel rays (Figure 40). Sound detectors have parabolic surfaces. Sound waves are reflected upon striking the surface and are concentrated at the focus. The mirror of a reflecting telescope is in the form of a parabola. Parallel rays of light from a distant heavenly body strike the parabolic surface and, reflected from it, meet at the focus within the telescope tube.

The hyperbola. The hyperbola is the path of a point moving so that the distance to one fixed point minus the distance to another fixed point is always the same. The diagram in Figure 41 shows a hyperbola in which F_1 and F_2 , called the foci, are the fixed points. P_1 is a point on the hyperbola. P_1F_1 minus P_1F_2 equals A_1A_2 . P_2 is another point on the hyperbola. P_2F_1 minus P_2F_2 is also equal to A_1A_2 .

The hyperbola is applied to LORAN, or long-distance navigation by the use of radar. We can give a general explanation of Loran by referring again to Figure 41. There are two radar stations, one called the master station (F_2 in the diagram) and the other the slave station (F_1). They are located on land about 300 kilometers apart. Electric pulsations are sent out from each station. It takes longer for such pulsations to travel from F_1 to P_1 than from F_2 to P_1 and from F_1 to P_2 than from F_2 to P_2 . The difference in time is a fixed constant for all points on the hyperbola. If the difference in time is greater or less, we have a different hyperbola.



At a very young age children are taught to use and recognize the basic geometric forms found in the world around them.

An airplane or surface vessel has a radar reception instrument which picks up this difference in time between the pulsations from the two stations F_1 and F_2 . The navigator consults a map upon which are drawn the various hyperbolas corresponding to the various differences in time. He consults the map and locates his own craft on the hyperbola. The hyperbola passing through the "home port" is then picked out, and the difference in the time of the pulsations is noted. The course is changed, until the radar receiver indicates that the craft is on the "home-port" hyperbola. The difference in pulsations is kept constant and the craft sails home along the hyperbola.

SOLID GEOMETRY

by Howard F. Fehr

The two-dimensional world of plane geometry does not suffice to explain the world in which we live: it is a world of three dimensions. In it, there are many planes, which are boundless and extend in every conceivable direction. There are also many kinds of curved surfaces. We must consider not only north, south, east, and west but also up and down. To explain this three-dimensional world, the branch of mathematics called solid geometry has been developed. We use this kind of geometry in building machines, skyscrapers, airplanes, steamships, bridges, and automobiles and also in explaining the phenomena of the universe.

In solid geometry, there are many more possible relationships between geometric elements than in plane geometry. In a single plane, two lines are either always parallel or else they intersect. In solid geometry, too, two lines may be parallel or else they may intersect, but they may also be *skew lines*. Skew lines are not in the same plane, are never parallel, and never intersect. Only one line, in a plane, can be drawn perpendicular to another line at a given point. In solid geometry, any number of such perpendicular lines can be drawn. For example, the spokes of a wheel are lines every one of which is perpendicular to the axle at the same point. In a plane, all points at a fixed distance from a fixed point are on a circle. In three-dimensional space, however, they are on a sphere containing

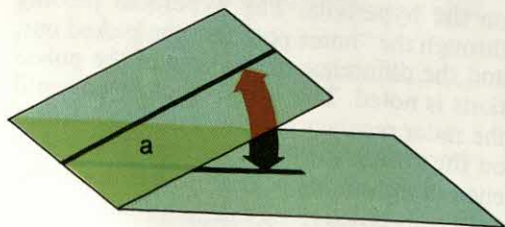
an infinite number of circles passing through the center.

ANGLES IN SOLID GEOMETRY

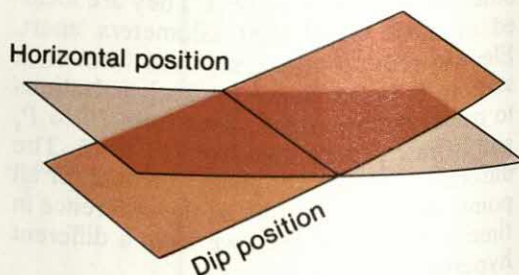
The simplest angle in solid geometry is called a *dihedral* ("two-faced") *angle*. It is formed by two intersecting planes. The size of this angle is measured by the *plane angle*. This is formed by two lines, one in each face, meeting the edge, or intersection of the two planes, at right angles (Figure 1). When an airplane banks its wings, the angle of bank is a dihedral angle between the horizontal and dip position of the wings (Figure 2). The dihedral angle is measured by an instrument in the plane, and the size of this angle determines in part the speed with which the airplane will change its direction of travel.

When three planes meet at a point, they form a *trihedral angle* (Figure 3). Each of the angles making up a trihedral angle is called a *face angle*. In Figure 3, ADC , CDB , and ADB are all face angles. If more than three planes meet in a point, the angle is called a *polyhedral* (many-faced) *angle*. The sum of the face angles of a polyhedral angle must be less than 360° . As the sum of the angles gets closer to 360° , the angle becomes less pointed until at 360° , it becomes a plane (Figure 4). The crystals of minerals show many kinds of polyhedral angles. An analysis of these

1. The intersecting planes below form a dihedral angle. a is the plane angle—the angle between the planes.



2. The dihedral angle between the horizontal and dip positions of an aircraft's wing is the angle of bank.



angles makes it possible to identify the various minerals.

FIVE COMMON SOLIDS

The major part of the study of solid geometry is based on five common solids: the prism, the cylinder, the pyramid, the cone, and the sphere.

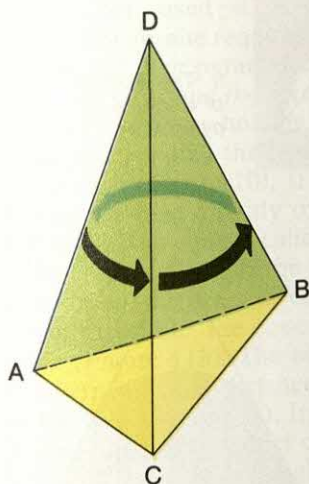
The prism. In a prism, all of the sides, or side faces, are parallelograms; the bases are parallel and equal polygons. Several kinds of prisms are shown in Figure 5. The most common of all is the cube, in which all the faces are square (Figure 5b). In a triangular prism (Figure 5a), the bases are triangles. Triangular prisms made of glass are used in studying the refraction, or bending, of light.

The most important properties of a prism are its area, or surface measure, and its volume, or space-filling measure. The lateral area, or area of the sides, is equal to the perimeter, or length of the boundary, of the base times the height. Of course, to find the total area of a prism, one adds the areas of the two bases to the lateral areas. The interior of a room is often a prism. If we want to decorate the walls, for example, we can find their total area by first finding the perimeter of the floor and then multiplying

it by the height of the room. If we want to paint both the walls and ceiling of a room, we find the lateral area of the room and then the area of the floor (which, in the ordinary room, is the same as the area of the ceiling). Then we add the two. It is by making such calculations that a painter or decorator estimates the cost of the materials needed in the work.

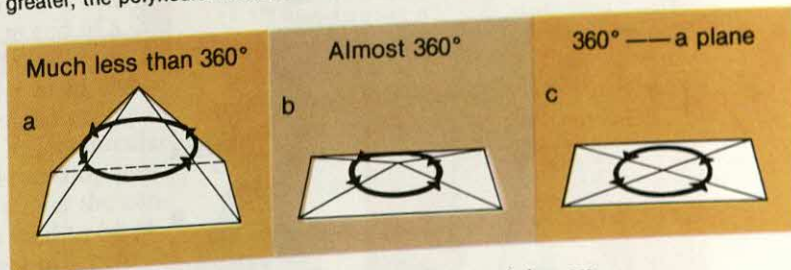
The volume of a prism is found by multiplying the area of the base by the altitude. This calculation is very important in the building of a house. Most builders estimate the construction cost as so much per cubic unit. To estimate how much it will cost to build a home, you must first find the total volume of the prisms of which the house will consist. If the volume in question is 800 cubic meters and the builder gives an estimate of \$30 per cubic meter, the cost will be approximately $800 \times \$30$, or \$24,000.

The cylinder. If one rotates a rectangle completely about one of its sides, as in Figure 6, it will define (mark the boundaries of) the solid called a cylinder. An ordinary tin can is a good example of a cylinder—a right circular cylinder, in which the bases are circles and the sides are perpendicular to the bases. It is not the only kind of cylinder, however. The bases of cylinders can have elliptical shapes and the sides are not always at right angles at the bases.

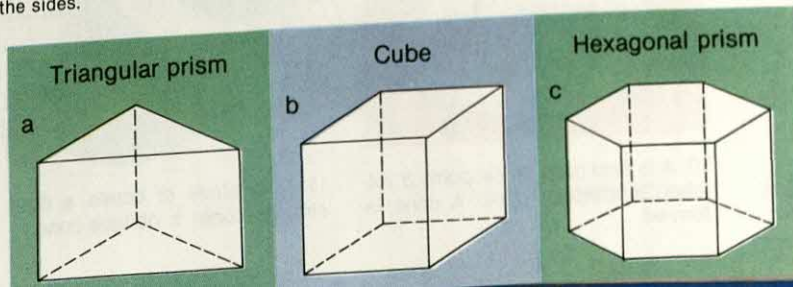


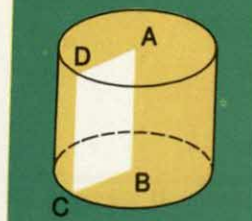
3. When three planes meet at a point, they form a trihedral angle, like the one shown above.

4. A series of polyhedral angles. As the sum of the face angles become greater, the polyhedral becomes less and less pointed. At 360° it is a plane.

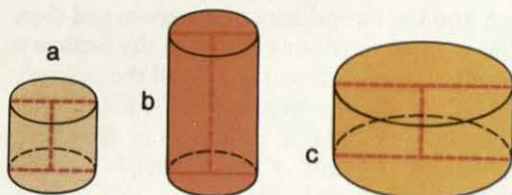


5. The triangular prism, cube, and hexagonal prism, shown below, are known as regular prisms. In all of them, the bases are at right angles to the sides.





6. If rectangle $ABCD$ is rotated about side AB , it will mark the boundaries of a cylinder, which will have AB as the axis.



7. If the height of cylinder a is doubled and the base remains the same, as in b , the volume is doubled. If the height remains the same and the diameter is doubled, as in c , the volume of the cylinder is increased fourfold.

The lateral area of a right circular cylinder is $2\pi rh$, in which π is approximately 3.1416, r is the radius of the base, and h is the height. Suppose that a canning company decides to manufacture a number of liter cans, and wants to know how much metal will be needed for the cans. The required quantity is found by adding the areas of the two circles that make up the bases to the area of the side.

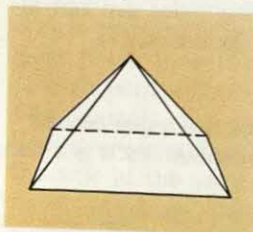
The volume of a cylinder is found by multiplying the area of the base by the height. The formula is $\pi r^2 h$, in which r is the radius of the base and h is the height. It is important to find the volume of a cylinder in computing the content of tin cans, gas-stor-

age tanks, reservoirs, and so on, and also in determining the rate of flow and pressure in pipes containing liquids.

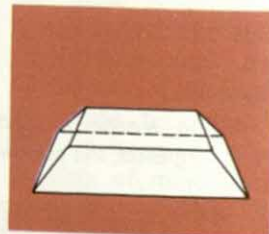
If the height of a cylinder is doubled, and the diameter remains the same, its volume will also be doubled. You can see that this is so by placing one can next to another just like it. If, however, the height of a cylinder remains the same and the diameter of the base is doubled, its volume will increase fourfold (Figure 7).

To see whether the economy size of a product sold in cans provides a real bargain, it would be very helpful to calculate the volume of the regular size and that of the economy size and then compare the two. Suppose the regular-size can is 12 centimeters tall and has a base with a diameter of 16 centimeters. Suppose the economy-size can is also 12 centimeters tall and has a base with a diameter of 20 centimeters. The regular size costs \$1.00 and the economy size \$1.25. The problem is: will we save money if we buy the economy size?

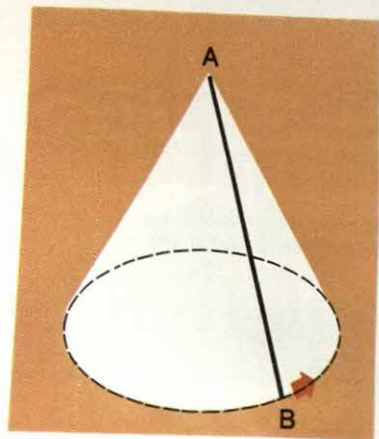
We know that the volume of a cylinder is $\pi r^2 h$. To find the volume of the regular can we substitute 8 for r (the radius is half the diameter) and 12 for h . Using 3.14 as the value of π , we would get $3.14 \times 64 \times 12 = 2,412$ cubic centimeters. This is the



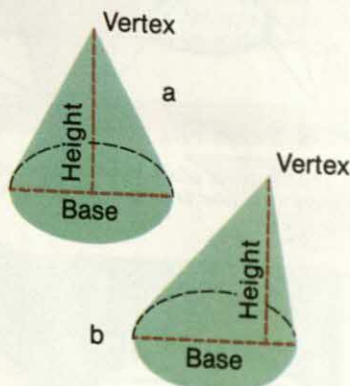
8. Square pyramid. It has a square base.



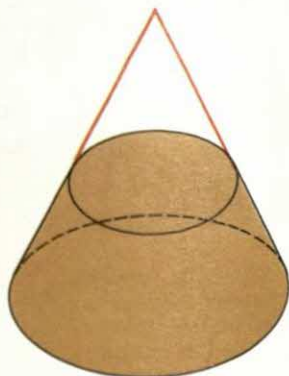
9. The frustum of a square pyramid.



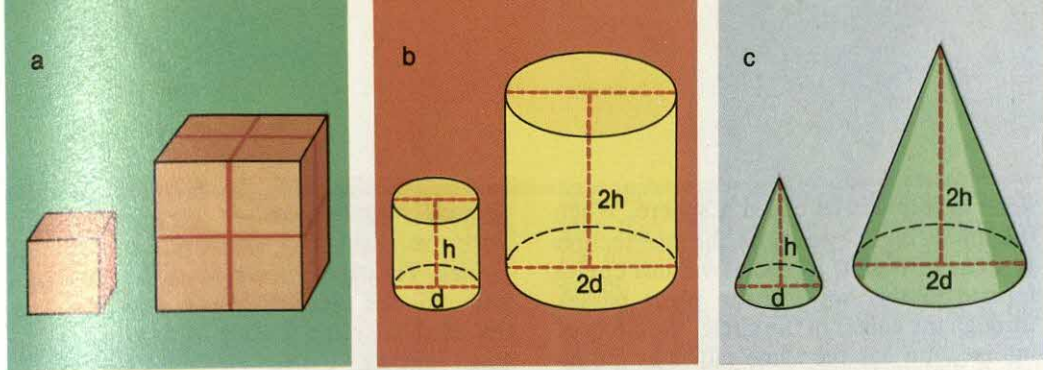
10. A is held fixed while point B follows a circular path. A cone is formed.



11. Two kinds of cones. a , right circular cone; b oblique cone.



12. Above is the frustum of a right circular cone.



13. Similar solids. Similar squares are shown in a; similar cylinders in b; similar cones, in c.

volume of the regular can. The volume of the economy-size can would be $3.14 \times 100 \times 12 = 3,768$ cubic centimeters. The economy-size can would hold about $1\frac{1}{2}$ times as much as the regular can ($3768 \div 2412$) and would cost $1\frac{1}{4}$ times as much. It would therefore represent quite a good bargain.

The pyramid and the frustum of a pyramid. In the solid called the pyramid, the sides are triangles whose vertices meet at a common point and whose bases form a plane (Figure 8). If the base is a square we have a square pyramid. The great pyramids of Egypt are square pyramids. So are the obelisks called "Cleopatra's needles."

When the top of a pyramid is cut off by a plane parallel to the base, the lower part is called a *frustum* (Figure 9). Army squad tents and coal hoppers, among other things, have the shape of the frustum of a pyramid. To calculate the volume of a frustum, a rather complex formula is required. Yet there is evidence that the ancient Egyptians had an exact formula for making such a calculation. They used it in determining the amount of granite required to build different sections of their pyramids.

The cone and the frustum of a cone. A cone is formed by holding one end of a line fixed and rotating the line, following a circular path (Figure 10). If the fixed end of the line is held directly over the center of the circle, the cone is called a right circular cone (Figure 11a). If the vertex (top-most part) of a cone is not directly over the center of the circle, the cone is said to be oblique (Figure 11b). The height of a cone is the perpendicular distance from the vertex to the base (Figure 11). Its volume is equal to one-third the product of the area of the base and the altitude (height).

If we divide a cone in two parts by passing through it a plane parallel to the base, the lower part is called a frustum

(Figure 12). Many machine parts are in the form of cones or frustums of cones.

CHARACTERISTICS OF SIMILAR SOLIDS

Solids which are of the same shape but of different sizes are said to be *similar*. The corresponding polyhedral angles of similar solids are equal, and the corresponding lines are in proportion.

The areas of similar solids have the same ratio as the squares of the corresponding linear parts (see Figure 13). In each of the sets of similar figures in Figure 13, the surface area of the larger figure is four times the surface area of the smaller, because the linear parts are twice as large. If the linear parts were enlarged three times, the area would be nine times as great.

The volume of similar solids has the same ratio as the cubes of the linear parts. In the sets of similar solids shown in Figure 13, the larger figure has eight times the contents of the smaller figure. In the case of the cubes in Figure 13a, you can count eight small cubes in the larger cube.

The ratio of areas and volumes of similar solids has many practical applications. All spheres are similar. If oranges 8 centimeters in diameter sell for 30 cents a dozen, while oranges of the same kind and 10 centimeters in diameter sell for 50 cents, which would be the better buy? The volume

of a large orange is $\left(\frac{10}{8}\right)^3$, or $\frac{1,000}{512}$, or about 2 times the volume of a small orange. Obviously in this case the large oranges would be a much better buy, since they cost less than twice as much as the smaller oranges but have twice the volume.

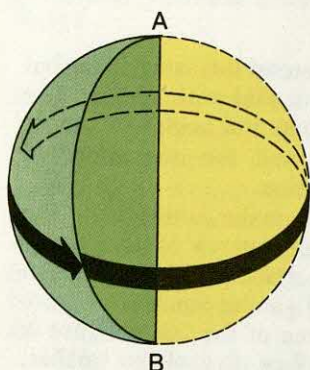
The sphere. If a semicircle is rotated about a diameter (Figure 14), the solid de-

fined in this way is called a sphere. When the sphere is cut by a plane, the intersection is a circle. Figure 15 shows various circles formed in this way. If the plane passes through the center of the circle, the circle of intersection has the same radius as the radius of the sphere. A circle such as this is called a great circle. All the other circles are small circles.

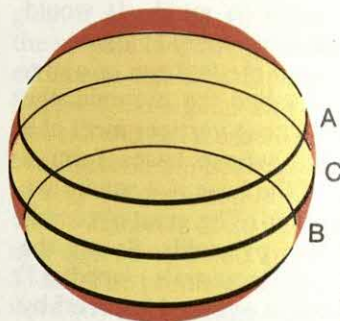
The earth may be considered as a sphere in which the north and south poles are the ends of a diameter called the axis

(Figure 16). The circles passing through both the north and south poles are great circles. They are known as circles of longitude. All the planes (except one) that cut the earth at right angles to the axis form small circles, called circles of latitude. There is just one plane that passes through the center of the earth and at right angles to the axis. It forms a great circle called the equator.

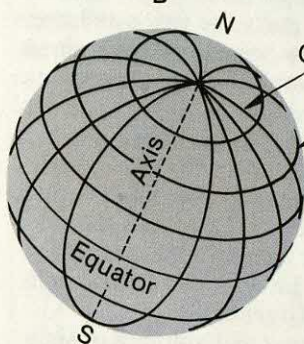
Between any two points on the earth (not including the poles) only one great cir-



14. A spherical surface is formed as the semi-circle is rotated around AB, which is the diameter.

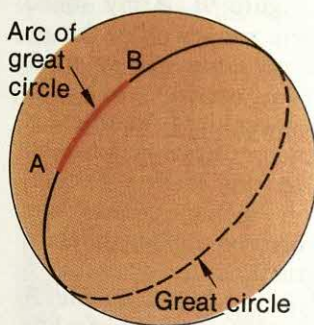


15. Plane C, passing through the center of the circle, is a great circle. A and B are small circles.



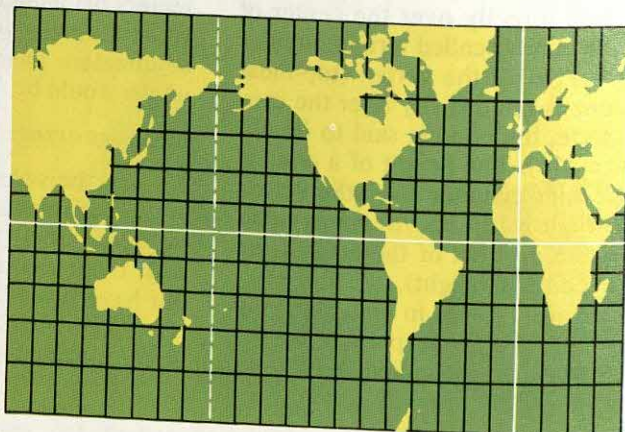
Circle of longitude
Circle of latitude

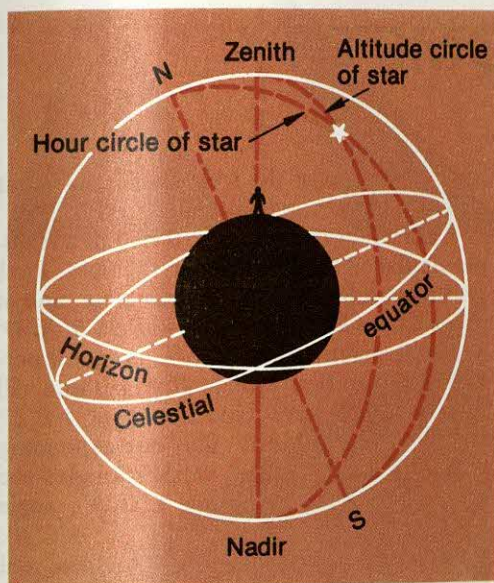
16. The earth as a sphere. The north and south poles are the ends of the axis, which is a diameter.



17. The shortest distance between A and B on the sphere is arc AB, forming part of a great circle.

18. The Mercator projection. On such a map, any place on earth can be located by determining the latitude and longitude of the place.





19. A simplified diagram showing the celestial sphere. Earth is the center of this sphere.

cle can be drawn (Figure 17). All other circles passing through the two points will be small circles. The shortest of all the arcs between the two points is the arc of the great circle (AB , in Figure 17). Pilots of planes, as far as possible, steer a course determined by the arc of a great circle between their starting point and destination.

If we think of the earth as a rubber ball and cut this ball along one half of a circle of longitude, we can stretch the ball to form a flat rectangular sheet. The circles of longitude will then become parallel vertical lines and the circles of latitude parallel and equal horizontal lines (Figure 18). This sheet now represents a rectangular map of the world. A map of this type is called a Mercator projection, after the 16th-century Flemish geographer Gerardus (or Gerhardus) Mercator, who developed it. The great longitudinal circle passing through Greenwich, England, is given the value 0° longitude. The equator is given the value 0° latitude. On such a map any place on earth can be located by determining the longitude and latitude. Also, on such a map, the farther we go from the equator, the more we find the original area stretched, so that land areas

near the poles seem much larger on the map than they really are on the earth. Users of the map must take such distortions into account.

The area of a sphere is four times the area of a great circle; the formula is $4\pi r^2$. The earth's radius is approximately 6,380 kilometers; hence the total surface of the earth is $4 \times \pi \times 6,380^2$ square kilometers, or about 511,000,000 square kilometers.

The volume of a sphere can be expressed by the formula $\frac{4}{3}\pi r^3$. Since the radius of the earth is 6,380 kilometers, the volume of our planet is $\frac{4}{3} \times \pi \times 6,380^3$, which gives about 1,090,000,000,000 cubic kilometers.

USE IN ASTRONOMY

Solid geometry has enabled astronomers to give a useful interpretation of the heavens and to calculate the distance and position of the celestial bodies. The universe is conceived of as a huge sphere with an infinitely great radius, which appears to revolve around the earth. In Figure 19, we give a greatly simplified presentation of such a sphere as seen from the vantage point of a man who is stationed at latitude 50° . This man stands on a much smaller sphere, which of course is the earth. Directly overhead is the zenith; directly below him is the nadir. The line where the sky seems to meet the earth is called the horizon. If the axis of the earth, which passes through the north and south poles, is extended, it will meet the outer bounds of our imaginary celestial sphere at the celestial poles—north and south. The line connecting the two celestial poles is the celestial axis. The plane of the earth's equator will cut the outer limits of the celestial sphere in a great circle called the celestial equator. A great circle passing through the poles and a star is the hour circle of the star. The altitude circle of the same star is a great circle passing through the zenith and the star.

These are but some of the features of the celestial sphere. They provide a frame of reference that enables the astronomer to trace the motions of celestial objects. This is one of the outstanding contributions of solid geometry to science.

TRIGONOMETRY

An important offshoot of geometry is trigonometry, or triangle measurement. In trigonometry, when certain parts of triangles are known, one can determine the remaining parts and thus solve a great variety of problems.

The founder of trigonometry was the Greek astronomer Hipparchus of Nicaea, who lived in the second century B.C. Hipparchus attempted to measure the size of the sun and moon and their distances from the earth. He felt the need for a type of mathematics that, by applying measurements made on the earth, would enable him to measure objects far out in space. He was led to the invention of trigonometry.

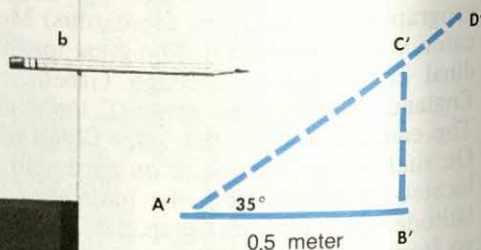
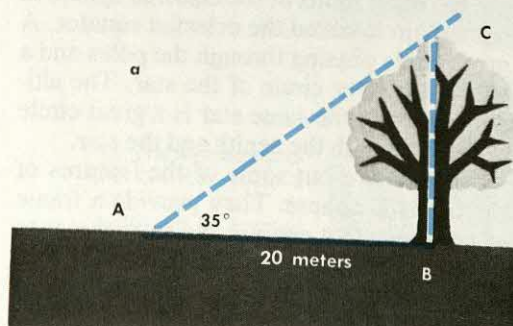
Boy scouts have occasion to use trigonometry in their field work. A common problem that is put to them is to find the height of a tree. The scout first measures a distance, say 20 meters, from the base of the tree, as shown in Figure 1. This will be his "base line." At A , he measures the angle from the ground to the treetop by means of a protractor. Let us suppose that this angle is 35° . The scout now knows two facts about the large right triangle formed when he connects points B (the base of the tree), A (the end of the line drawn from the tree),

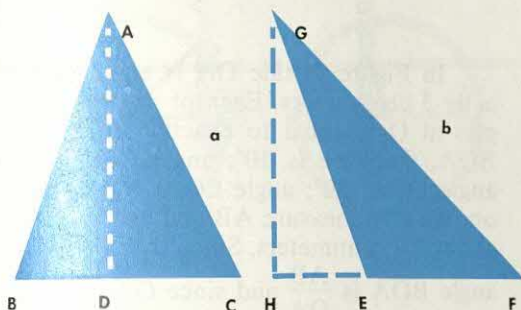
and C (the top of the tree). He knows that AB is 20 meters and that angle BAC is 35° .

On paper our scout now makes a triangle similar to the large one in the field. First he draws a line $A'B'$ $\frac{1}{2}$ meter long, and at A' he draws an angle of 35° —angle $B'A'D'$ —with a protractor. Next he erects a perpendicular to line $A'B'$ at B' . This line will intersect $A'D'$ at C' ; and the angle $A'B'C'$ will be a right angle. The corresponding angles of the large and small triangles are equal: angle CAB = angle $C'A'B'$; angle ABC = angle $A'B'C'$; angle ACB = angle $A'C'B'$. Hence we have two similar triangles, and the corresponding sides will be proportionate.

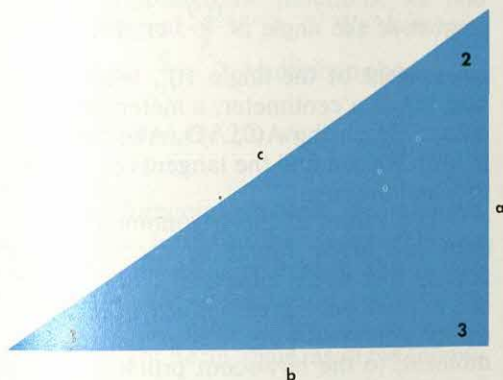
The scout now measures line $B'C'$ and finds that it is 35 centimeters, or 0.35 meter. Since the sides of the similar triangles are in the same ratio, AB is to $A'B'$ as BC is to $B'C'$. We know all these quantities except BC , which we can call x . We now have the proportion 20 is to 0.5 as x is to 0.35 which we can write as $20:0.5::x:0.35$. In any proportion, the product of the extremes (the two outer terms) is equal to the product of the means (the two inner terms); hence $x = 14$. The height of the tree, then, is 14 meters. The scout solved this problem knowing only one side and an acute angle of a right triangle. He used the basic methods of trigonometry, though a mathematician, as we shall see, would not go at the problem in that particular way.

1. The problem is to find the height of the tree when the angle at A (35°) and the distance AB (20 meters) are known. We show how to solve the problem by means of similar triangles ABC and $A'B'C'$.





2. Any triangle can be converted into a right triangle by drawing a perpendicular from the vertex to the base.



3. A typical right triangle with the hypotenuse labeled c , and legs a and b .

TRIGONOMETRIC FUNCTIONS

Trigonometry is based on the use of the right triangle. It can be applied to any triangle because by drawing an altitude (a perpendicular from the vertex to the base) we can always convert it into right triangles. In Figure 2a, for example, the altitude AD divides the triangle ABC into the right triangles ADB and ADC ; in Figure 2b, the altitude GH converts the triangle GEF into right triangles GHF and GHE .

Certain basic ratios or relationships between the sides of a right triangle are the very heart of the study of trigonometry. Among these ratios are the sine, cosine, tangent, and cotangent. To understand what these terms mean, let us draw a typical right triangle with angles 1, 2, and 3 and sides a , b , and c (Figure 3). Angle 3 is a right angle; the other two angles are acute angles (that is, angles of less than 90°). Side c , which is opposite the right angle, is the hypotenuse. The other two sides, a and b , are called legs. We can now define sine,

cosine, tangent, and cotangent as follows:

The *sine* of either of the acute angles is the ratio of the opposite leg to the hypotenuse. The sine of angle 1 is a/c ; the sine of angle 2 is b/c .

The *cosine* of either of the acute angles is the ratio of the adjacent leg to the hypotenuse. The cosine of angle 1 is b/c ; the cosine of angle 2 is a/c .

The *tangent* of either of the acute angles is the ratio of the opposite leg to the adjacent leg. The tangent of angle 1 is a/b ; the tangent of angle 2 is b/a .

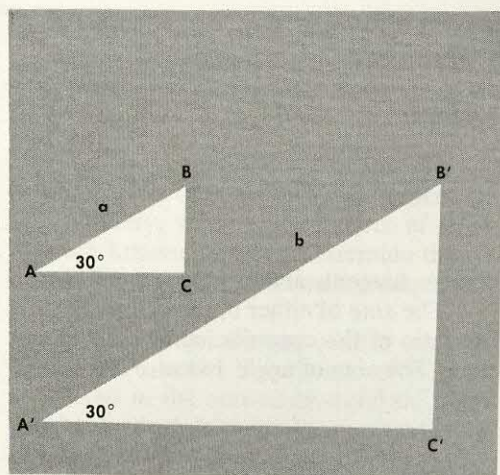
The *cotangent* of either of the acute angles is the ratio of the adjacent leg to the opposite leg. The cotangent of angle 1 is b/a ; the cotangent of angle 2 is a/b .

A sine of an angle is said to be a trigonometric function of that angle, because its value depends upon the size of the angle. The cosine, tangent, and cotangent are also trigonometric functions.

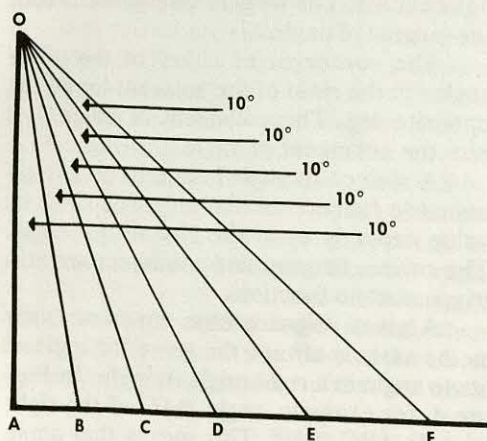
A given trigonometric function, such as the sine, is always the same for a given acute angle in a right-angle triangle. In Figure 4, for example, angle BAC of the right triangle ABC is 30° . This means that angle ABC must be 60° since angle ACB is 90° and the sum of the interior angles of a triangle is 180° . Angle $B'A'C'$ of the right triangle $A'B'C'$ is 30° and angle $A'B'C'$ must be 60° . Hence the corresponding angles of the two triangles are equal, and the corresponding sides must be in the same ratio. Since

this is so, $\frac{BC}{AB}$ (the sine of the 30° angle BAC) and $\frac{B'C'}{A'B'}$ (the sine of the 30° angle

$B'A'C'$) must be equal. Hence the sine of 30° is always the same no matter how large or how small the right triangle in which it occurs.



4. The sine of angle BAC (30°) is equal to the sine of angle B'A'C' in the two similar triangles.



5. This diagram shows how mathematicians have determined the tangents of angles.

PROBLEM SOLVING

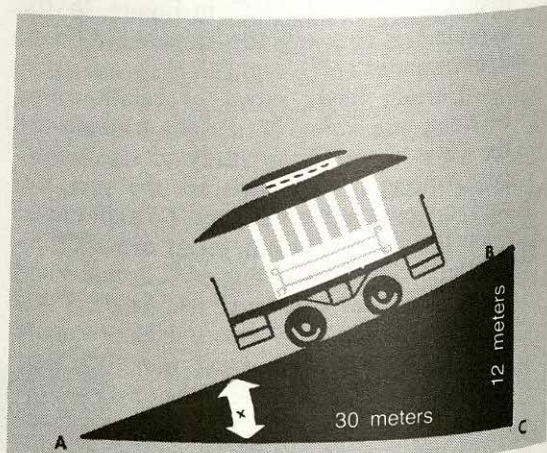
Mathematicians have worked out the values of the trigonometric functions. By way of example, the sine of 40° , to five decimal places, is .64279; its cosine, .76604; its tangent, .83910; its cotangent, 1.1918. To give some idea of how such figures are derived, let us examine the procedure for finding out the tangents of different acute angles. You will recall that the tangent of an angle is the ratio of the opposite leg to the adjacent leg.

In Figure 5, side OA is equal to exactly 5 centimeters. Each of the small angles at O is equal to exactly 10° . Angle BOA, therefore is 10° ; angle COA, 20° ; angle DOA, 30° ; angle EOA, 40° ; and so on. We now measure AB and find that it is about 0.9 centimeters. Since the tangent of angle BOA is $\frac{AB}{OA}$ and since $OA = 5$, the

tangent of the angle is $\frac{0.9}{5}$, or .18. This is the tangent of the angle 10° , whether the side OA is a centimeter, a meter, or a kilometer. Measuring AC, AD, AE, and so on in turn, we can find the tangents of 20° , 30° , 40° , and the rest.

The values of the trigonometric functions are to be found in special tables. Armed with these tables, it is possible for one to work out a great variety of measurements with great ease. Let us return, for a moment, to the boy-scout problem as presented in Figure 1a. We know that AB is 20 meters and that angle CAB is 35° . We are to find out BC (the height of the tree). The tangent of the 35° angle CAB is $\frac{BC}{AB}$. We know that AB is 20 meters; since BC is the unknown quantity, we call it x . Hence the tangent of angle CAB is $\frac{x}{20}$. Looking up the

6. The cable car rises 12 meters in a horizontal distance of 30 meters. What is the angle at A?



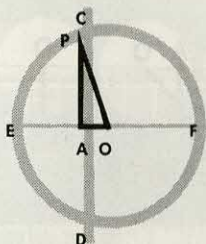
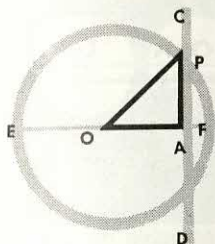


table of trigonometric functions, we find that the tangent of 35° is .7. We then have the equation $\frac{x}{20} = .7$. Multiplying both sides of the equation by 20, we have $\frac{20x}{20} = 14$. $x = 14$.

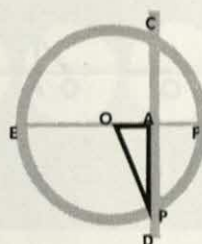
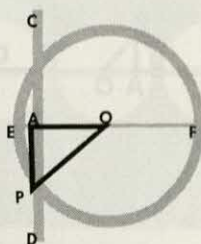
In the foregoing problem, the unknown quantity was a part of the tangent ratio. In other trigonometric problems, the cotangent or the sine or the cosine might be involved. In still other cases, the unknown quantity might be an angle, as in the following problem.

A cable car going up a hill in a uniform slope rises 12 meters in a horizontal distance of 30 meters. What is the angle of the slope to the nearest degree? We diagram the problem as in Figure 6. We want to find angle x . We know that the tangent of x is $\frac{BC}{AC} = \frac{12}{30} = .4$. Consulting the table of trigonometric functions, we note that .4 is the tangent of the angle 22° (to the nearest degree). Therefore the angle of the slope is 22° .

The surveyor makes extensive use of trigonometry. First, the surveyor tries to get a fixed line that has no obstruction so that it can be measured fairly accurately. This is the base line that is used in all calculations. Other measurements, as far as possible, are measurements of angles. Trigonometry is also of vital importance in engineering, navigation, mapping, and astronomy.

USED TO STUDY PERIODIC PHENOMENA

Trigonometry is used in still other ways. For one thing it serves in the study of various periodic phenomena. Any phenomenon that repeats itself in regular intervals of time is called periodic. The tides, for

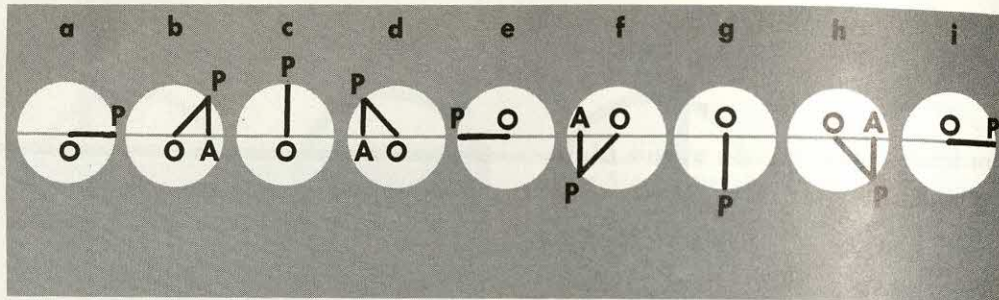


7. As radius OP goes around the circle in a counterclockwise direction, it makes a series of triangles with the diameter of the circle and with the rod CD kept in a vertical position at the end of the radius.

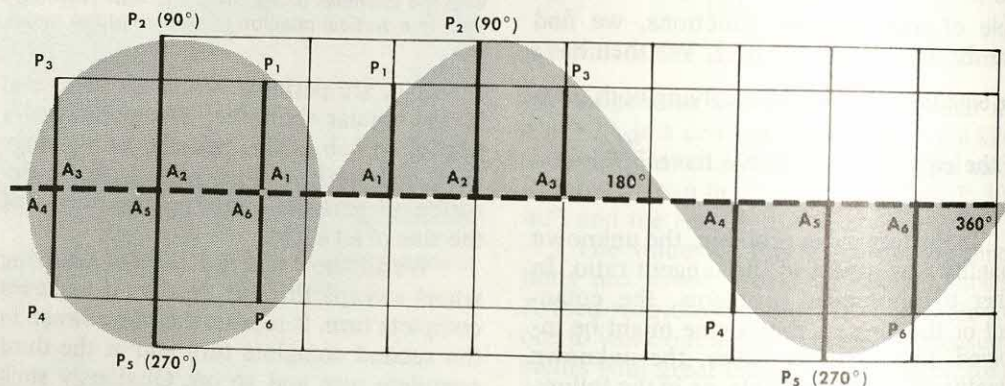
example, are periodic, since they rise and fall in regular sequence. The motion of a pendulum bob is also periodic as it swings back and forth. Let us show how we describe all periodic phenomena in terms of the sine of an angle.

We all know that the spoke of a moving wheel sweeps through 360° as it makes a complete turn. It repeats the same sweep in the second complete turn and in the third complete turn and so on. Obviously such rotation is periodic. The spoke of a wheel is really the radius of a circle. We can analyze the motion of the radius around the center of the circle by examining the diagrams in Figure 7. We are to suppose that a rod, CD , is kept in a vertical position at the end of the radius as the latter moves around the circle. You will note that a series of right angles is formed as the rod maintains its vertical position. The line AP , joining the points where CD meets the circumference of the circle and the diameter EF , grows longer and then shorter. Angle POA , which is called the angle of rotation, also changes as the radius goes around the center of the circle. The sine of POA , the angle of rotation, is $\frac{AP}{OP}$. OP , the hypotenuse of the right triangle APO , is the radius of the circle and of course never changes. If we give it the value unity (that is, 1), AP will represent the sine of the angle of rotation, since the sine is equal to $\frac{AP}{OP}$.

Let us now analyze the different sine values of the angle of rotation as the radius sweeps around the circle. Sine values above the diameter are expressed as positive values. Values below the diameter are negative values. In Figure 8a, the sine is



8. As the radius sweeps around the circle, the sine values of the angle of rotation vary.



9. Variations in sine value are shown here by means of a graph. The different values of the sine PA are given here as P_1A_1 , P_2A_2 , P_3A_3 , and so on. The curved line at the right of the circle is known as a sine curve.

zero, corresponding to a zero angle of rotation. The sine continues to grow as the radius revolves around the center (b) until it reaches the value 1 in c . It becomes smaller (d) until it reaches zero again in e . Then it goes below the diameter and is assigned negative values. It increases in size (f) until it reaches the value -1 in g . It becomes smaller thereafter (h) until, after a complete rotation has taken place, it becomes zero again (i).

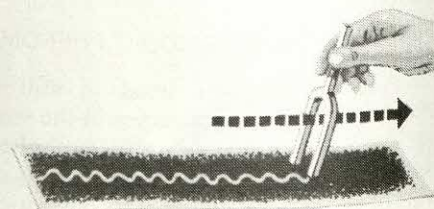
We can show the variation in the sine by the line graph in Figure 9. We give the different values of the sine, PA , as P_1A_1 , P_2A_2 , P_3A_3 , and so on. The line at the right of the circle is called the sine curve. It repeats itself every 360° .

The sine curve can be applied, among other things, to the periodic phenomena of sound. For example, when a tuning fork is struck, it vibrates and the vibration results in sound waves being sent out. If, immediately after being struck, a tuning fork which gives a tone of middle C is drawn very rapidly over a sheet of paper covered with

soot, the vibration of the fork will describe a series of sine curves (Figure 10). Two hundred sixty-four complete oscillations are produced in a second; hence middle C corresponds to 264 vibrations per second. We say that it has a frequency of 264 hertz. Suppose we strike a tuning fork that gives a tone one octave higher than middle C and draw it over the paper as before. In this case 528 sine curves will be produced in a second. That means that the frequency is twice as great as before.

Sine curves of various amplitudes and frequencies are used to explain phenomena of electricity, light (both polarized and plane), and force, as well as those of sound. Thus trigonometry helps to explain and control our physical environment.

10. Sine waves made by a vibrating tuning fork as it was passed rapidly over a sheet of paper covered with soot.



ANALYTIC GEOMETRY

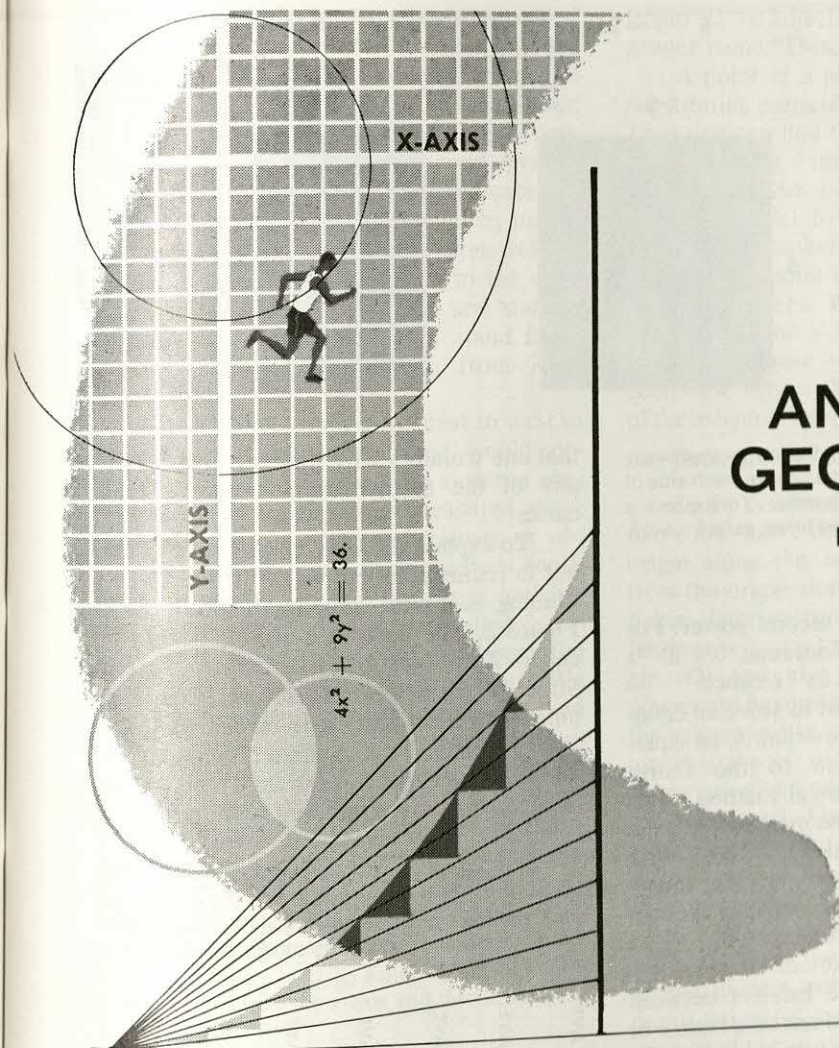
by Howard F. Fehr

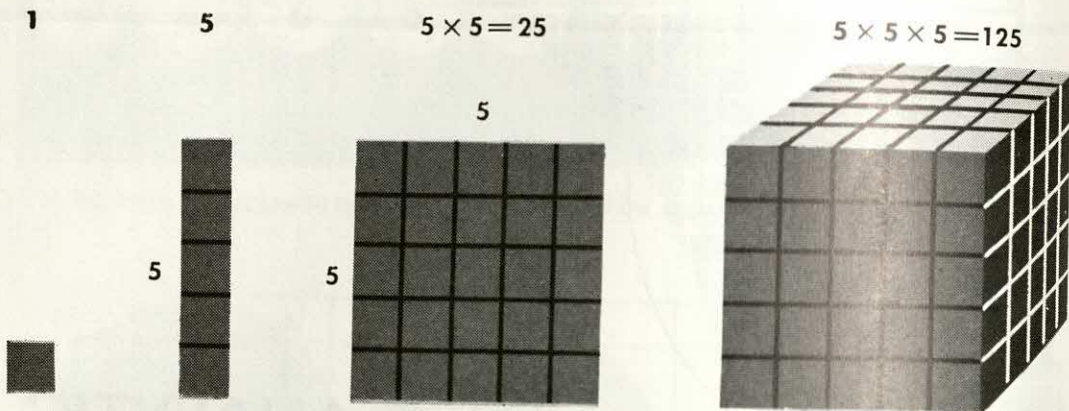
Analytic geometry explains geometric figures in terms of algebraic formulas. From the earliest times, algebra and geometry had been studied as separate subjects. No one had seriously considered before the eleventh century A.D. that numbers could be used to represent a point or line, or that a geometric figure could serve to represent the value of a number.

One of the first to try to combine algebra and geometry was the 12th century Persian poet and mathematician Omar Khayyam. He wrote a work on algebra which was clearly influenced by earlier Arab and Greek writings. In this work, Omar

showed how to solve algebraic equations by the use of squares, rectangles, and cubes. For example, for a number multiplied by itself, as in $x \times x$, he would use a square, each side of which had a length equal to the value (x). If x were equal to 5, each side of the square would be 5 units long, and the square would be made up of 25 units. If a number were to be a factor three times, as in $x \times x \times x$, Omar would use a cube, each side of which had a length equal to x . See Figure 1.

It was because of this method of solving equations that a number multiplied by itself came to be known as the square of the





1. To express a number multiplied by itself—say 5×5 —Omar Khayyam used a square, each side of which was equal to that number. To express a number used as a factor three times, as in $5 \times 5 \times 5$, he used a cube.

number, instead of the second power. For the same reason, we do not read “ x^3 ” as “ x to the third power” but as “ x cubed.”

Omar made no effort to solve an equation of the fourth degree—that is, an equation containing a term to the fourth power—by the geometrical method. This was obviously because no mathematician in his day had even an inkling of the fourth dimension. (Today, however, the fourth dimension is an accepted and vital concept in modern physics and mathematics.)

The pioneering effort of Omar Khayyam to break down the barriers between algebra and geometry did not bear fruit. For centuries afterward, the two subjects were still studied in watertight compartments, so to speak. It was not until the seventeenth century that the partition between these compartments was knocked down and that analytic geometry was finally developed.

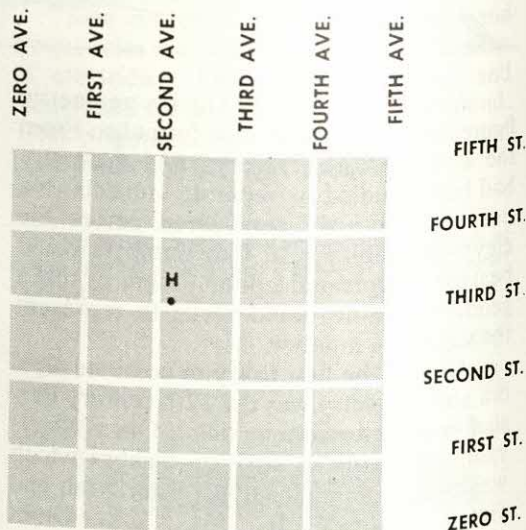
CARTESIAN CO-ORDINATE SYSTEM

The fundamental idea of analytic geometry was worked out by the great 17th century French philosopher and mathematician René Descartes, who claimed that the idea came to him in a dream. He presented this new approach to mathematics in his *Discourse on Method*, in 1637. The basic idea of analytic geometry is so simple

that one wonders why it never occurred to any of the mathematicians before Descartes.

To explain the idea, think of a city laid out in rectangular blocks, with the streets running east-west and north-south, as in Figure 2. The east-west streets are to be called Zero Street, First Street, Second Street, Third Street, and so on. Those running north and south are to be designated as Zero Avenue, First Avenue, Second Avenue, Third Avenue, and so on. If the houses in each block are numbered by hundreds, it

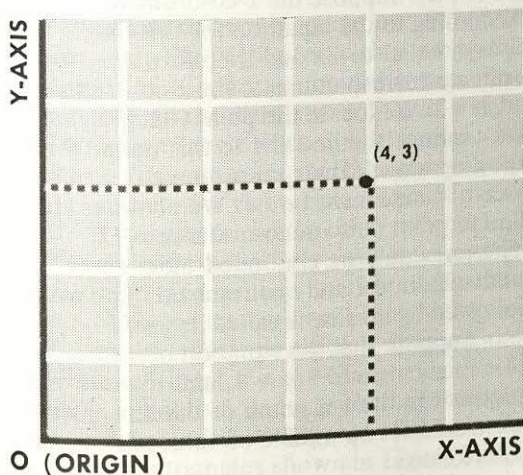
2. The house that we have indicated by *H* in this drawing is on 210 Third Street. It is 2.1 units to the east, starting from Zero Avenue and counting each block as a unit. It is 3 units to the north, starting from Zero Street.



will be easy to locate any house provided we know the name of the street and the street number. For example, if there is a house (H in Figure 2) at 210 Third Street, we know that it is on Third Street, one-tenth of the distance from Second Avenue to Third Avenue. We could indicate the location of the house more simply by means of two numbers, 2.1 and 3, written as $(2.1, 3)$. 2.1 would mean 2.1 units to the east, counting each block as a unit and starting from Zero Avenue; 3 would stand for 3 units to the north, starting from Zero Street.

Zero Street, going from east to west in the city we have just described, would correspond to what is called the x -axis in analytic geometry. Distances measured along this axis are known as x -distances, or *abscissas*. Zero Avenue, running from south to north, corresponds to the y -axis, and distances measured along it are called y -distances, or *ordinates*. The intersection of the x -axis and the y -axis is called the *origin*. We show the x -axis, y -axis and origin (marked O) in Figure 3.

A point in a plane can be located by an ordered pair of numbers (x, y) , with the x always coming first. x stands for the number of units measured along the x -axis from the origin; y , for the number of units measured along the y -axis from the origin. x and y are called *Cartesian co-ordinates*. "Cartesian" is the adjective corresponding to the proper name "Descartes." To locate point $(4, 3)$, go 4 units to the right from O , the origin, along the x -axis and then 3 units up, parallel to the y -axis.



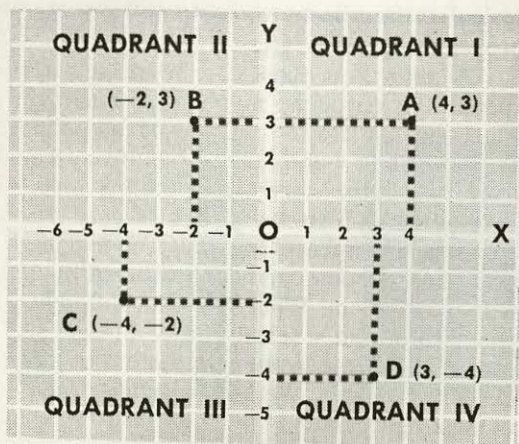
tesian" is the adjective corresponding to the proper name "Descartes."

A point in a plane can be located by substituting particular numbers for x and y . Thus one can find the position of the point $(4, 3)$ by going 4 units to the right from the origin along the x -axis, and then 3 units upward parallel to the y -axis (Figure 3). Any pair of numbers, such as $(4, 3)$ locates a point. Any point can be represented by a pair of numbers. What is more, different values of x and y select points all over a plane, and these points describe various geometric figures. We shall examine some of these figures later.

EXTENDING THE SYSTEM

In the system developed by Descartes, one could move only to the right from the origin along the x -axis and only upward from the origin along the y -axis. Mathematicians later extended the system by continuing the x -axis to the left beyond the origin (O) and also continuing the y -axis downward beyond O , as in Figure 4. This is the system in use at the present time. Distances to the left of the origin and below the origin are given negative values, as indicated in Figure 4. The four parts into which a plane is divided by the x -axis and y -axis are called *quadrants*. They are indicated by Roman numerals as shown. In the original system of Descartes, a point had to be somewhere in the first quadrant. In the present system, points corresponding to different values of x and y may be located in any one of the quadrants.

Of course, one or both of the two reference numbers, or co-ordinates, of a point may be negative. There are four different possibilities. (1) Both x and y may be positive; (2) both x and y may be negative; (3) x may be positive and y negative; (4) x may be negative and y positive. Let us consider some typical examples (Figure 4). Suppose the co-ordinates are $(4, 3)$. Moving 4 units to the right from O , the origin, along the x -axis, and then 3 units upward, parallel to the y -axis, we locate the point at A , in quadrant I. The second pair of co-ordinates we shall consider is $(-2, 3)$. We go 2 units to the left from O along the x -axis and then 3



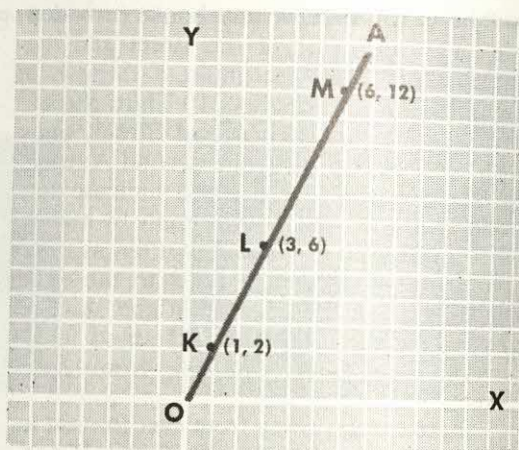
4. Each of the points $(4, 3)$, $(-2, 3)$, $(-4, -2)$ and $(3, -4)$ is located in a different quadrant.

units upward. We locate the point at B , in quadrant II. If the co-ordinates of a point are $(-4, -2)$, we go 4 units to the left from O along the x -axis and 2 units down, locating the point at C , in quadrant III. Finally, if the co-ordinates are $(3, -4)$, we go 3 units to the right from O along the x -axis and then 4 units down. The point is located at D , in quadrant IV.

The Descartes system of co-ordinates, as extended by later mathematicians, is particularly valuable because it enables us to analyze the geometric figure described by a variable point. The point has a series of different co-ordinates as it is selected for different positions. Under certain conditions, we can write an equation that will hold true for all possible positions of such a point. The equation can then be used in place of the geometric figure consisting of all the points. There are equations for straight lines, for circles, for ellipses, for parabolas, for hyperbolas.

GRAPH OF A LINE

Let us first consider the equation for a straight line, OA , which passes through the origin, O (Figure 5), and then through points K , L , and M . The co-ordinates of point K are $(1, 2)$. We locate it by going 1 unit to the right from O and 2 units upward. Point L has the co-ordinates $(3, 6)$; that is, to locate it, we go 3 units to the right from



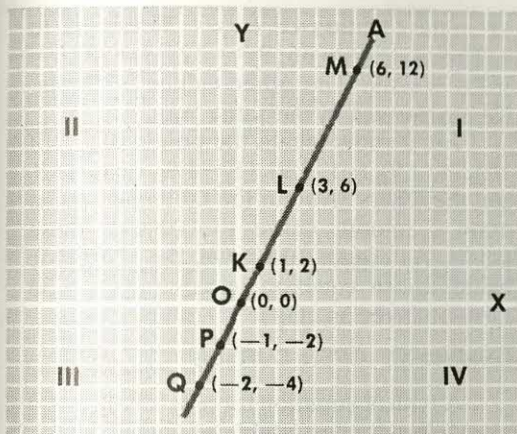
5. The equation for this straight line is $y = 2x$. The y -coordinate here is always twice the x -coordinate.

O and 6 units upward. The co-ordinates for M are $(6, 12)$. You will note that in each instance the y -co-ordinate is twice the x -co-ordinate. Hence the equation that expresses the relation between x and y for this particular line is $y = 2x$.

(We could also give it as $x = \frac{y}{2}$.)

All the co-ordinate values we have presented thus far for the line OA are positive values, and the points whose co-ordinates are $(1, 2)$, $(3, 6)$ and $(6, 12)$ are all in quadrant I. Line OA could be extended to quadrant III, as shown in Figure 6. As we have pointed out, the line passes through the origin, O . At this point, the co-ordinates are $(0, 0)$. Suppose the x -co-ordinate is -1 . According to the equation for this line, y is always equal to $2x$ and therefore the y -co-ordinate corresponding to the x -co-ordinate -1 is -2 . We locate the point $(-1, -2)$, and we continue the line OA to this point (P in the diagram). If the x -co-ordinate is -2 , the y -co-ordinate must be -4 . We continue the line to point Q , with co-ordinates $(-2, -4)$. The line could be continued indefinitely in both quadrant I and quadrant III. The path followed by the line is called its *graph*.

Suppose that the equation of a straight line is given as $x - y = 1$, and that we are required to draw a graph of the line. First, we would set up a table of values for x and y , as follows:



6. Line OA, located in Figure 5, is extended into quadrant III by locating points P (-1, -2) and Q (-2, -4).

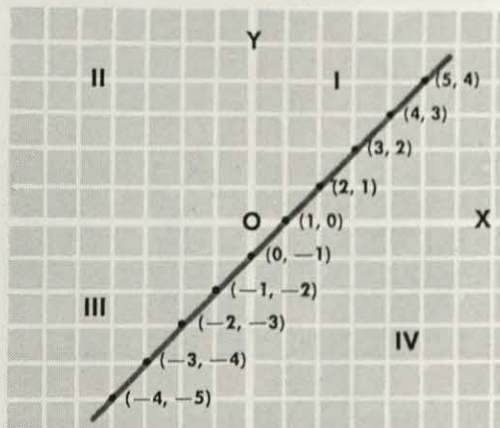
| | | | | | | | | | | | |
|---|---|---|---|---|---|----|----|----|----|----|----|
| x | 5 | 4 | 3 | 2 | 1 | 0 | -1 | -2 | -3 | -4 | -5 |
| y | 4 | 3 | 2 | 1 | 0 | -1 | -2 | -3 | -4 | -5 | -6 |

The y values corresponding to the x values in the table can be derived easily enough since in every case the y value is the x value minus 1. (Remember that the equation of this line is $x - y = 1$.) Once we have set down the x and y values, we locate the points (5, 4), (4, 3), (3, 2), (2, 1) and so on, as indicated in Figure 7, and we draw the line connecting the points. This is the straight line whose location is indicated by the equation $x - y = 1$. Note that the line passes through quadrants I, IV, and III.

GRAPH OF A CIRCLE

We can derive the equation of a circle if we apply the Pythagorean theorem to the system of Cartesian co-ordinates. The Pythagorean theorem states that in a right triangle (a triangle with a right angle), the sum of the square on the hypotenuse (the side opposite the right angle) is equal to the sum of the squares on the other two sides.

Let us draw a circle with its center at the origin and with a radius of 5 units (Figure 8). No matter what point on the circle we select, the x-distance, y-distance and radius will form a right triangle, with the radius, equal to 5 units, as the hypotenuse. According to the Pythagorean theorem, in each of the triangles shown in Figure 8, the



7. A straight line whose equation is $x - y = 1$. Note that this line extends through three quadrants.

square of the x-distance plus the square of the y-distance will be equal to the square of the hypotenuse—that is, 5^2 or 25. In other words, $x^2 + y^2 = 25$; and that is the equation of the circle shown in Figure 8.

Suppose we want to see whether a given point is on the circumference of this circle. If the sum of the squares of the point's co-ordinates is equal to 25, the point is on the circle. Take the point (3, 4). The sum of 3^2 and 4^2 is 25; hence (3, 4) is on the circle. So is (-4, -3), since the sum of $(-4)^2$ and $(-3)^2$ is 25. But point (4, 5) is not on the circle, since the sum of 4^2 , or 16, and 5^2 , or 25, is not equal to 25 (Figure 9).

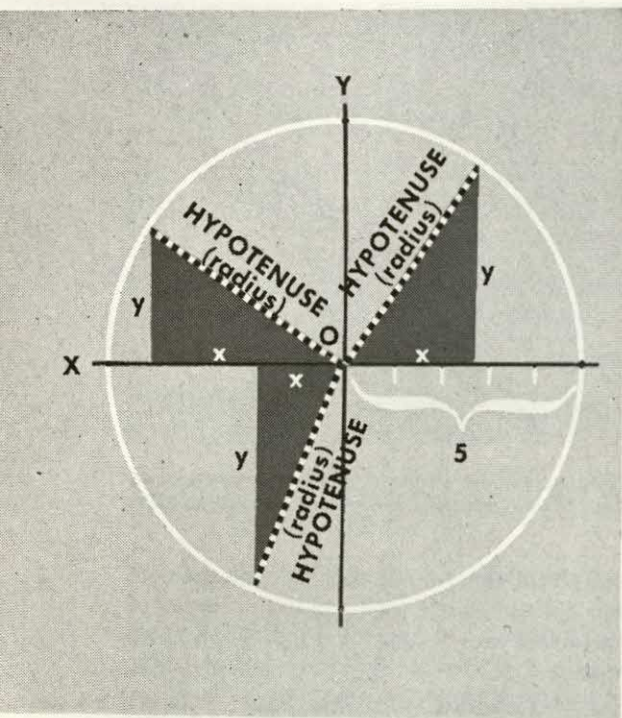
GRAPH OF AN ELLIPSE

Suppose that we are given the information that $4x^2 + 9y^2 = 36$ is the equation of a geometric figure. Let us prepare a graph and see what sort of a figure it is. First, we set up a table of values:

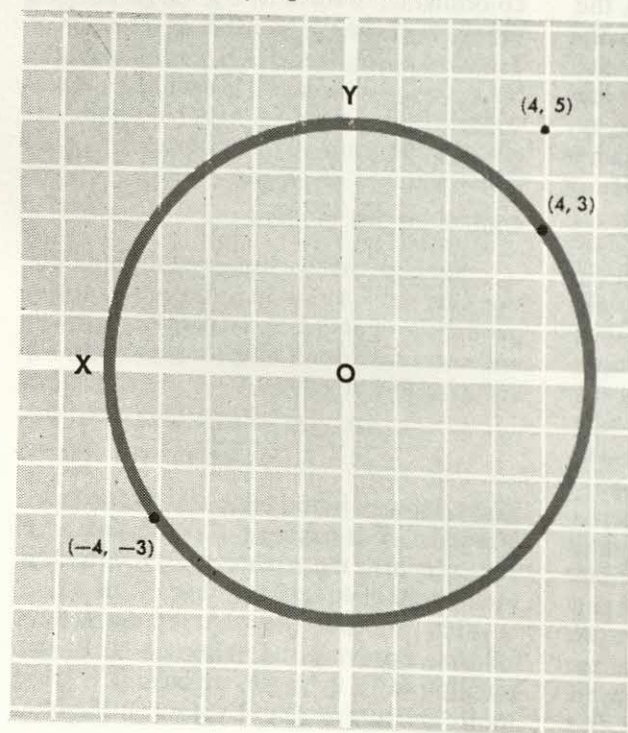
| | | | | |
|---|---------|-----------|-----------|---------|
| x | 0 | ± 1 | ± 2 | ± 3 |
| y | ± 2 | ± 1.9 | ± 1.5 | 0 |

The symbol \pm in the table stands for "plus or minus." For example, 4 is the square of + 2; it is also the square of - 2. Hence a square root of 4 is either + 2 or -2, or ± 2 .

The values in the table are derived as follows. Applying the value $x = 0$ to the equation $4x^2 + 9y^2 = 36$, we have $0 + 9y^2 = 36$; $9y^2 = 36$; $y^2 = 4$; $y = \sqrt{4} = \pm 2$. If x is 1,



8. The equation of the above circle is $x^2 + y^2 = 25$. The text explains how to derive the equation of a circle by applying the Pythagorean theorem.



9. The equation of the circle at left is $x^2 + y^2 = 25$. Point (4,3) is on the circle, since $4^2 + 3^2 = 25$. Point (-4, -3) is also on the circle. But point (4,5) is not on this particular circle, because $4^2 + 5^2$ is not equal to 25.

$4x^2$ in the equation $4x^2 + 9y^2 = 36$ is equal to 4. Then $4 + 9y^2 = 36$; $9y^2 = 32$; $y^2 = 3.6$; $y = \sqrt{3.6} = \pm 1.9$. The other values in the table are derived in the same way.

Let us now locate the points (0, 2), (0, -2), (1, 1.9), (-1, 1.9), (1, -1.9), (-1, -1.9) and the others given in the table, and let us draw a smooth curve to connect these points. As Figure 10 shows, the geometric figure whose equation is $4x^2 + 9y^2 = 36$ turns out to be an ellipse.

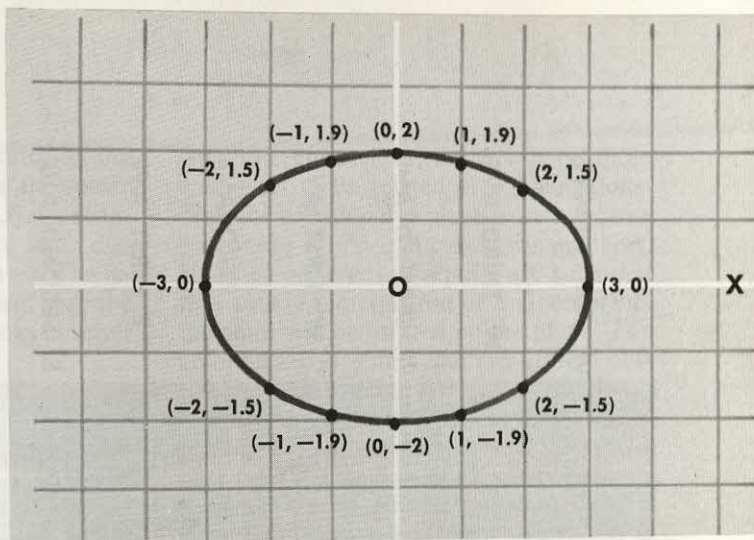
GRAPHS OF OPEN CURVES

A typical equation for a parabola is $x = y^2 - 4$. We prepare a table of values, as follows:

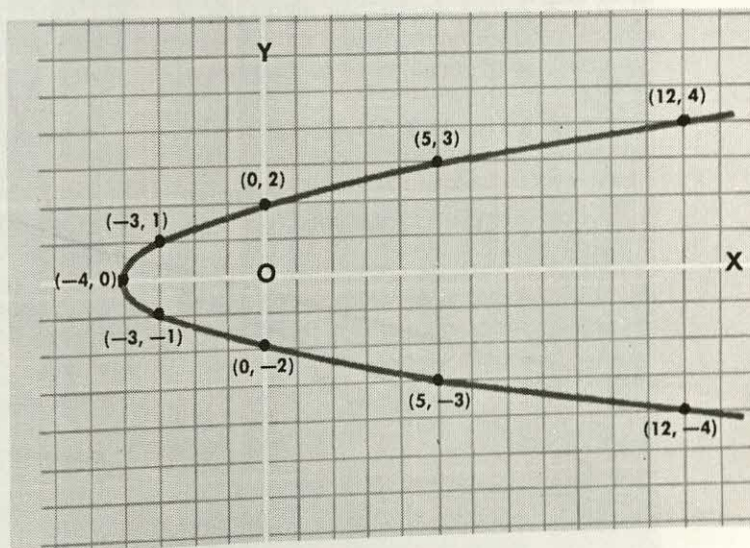
| | | | | | |
|-----|----|---------|---------|---------|---------|
| x | -4 | -3 | 0 | 5 | 12 |
| y | 0 | ± 1 | ± 2 | ± 3 | ± 4 |

If we plot the values (-4, 0), (-3, 1), (-3, -1), (0, 2), (0, -2) and the others given in the table, we have the parabola shown in Figure 11. Since the higher the values of x , the greater the corresponding values of y , the parabola is an open curve extending without limit.

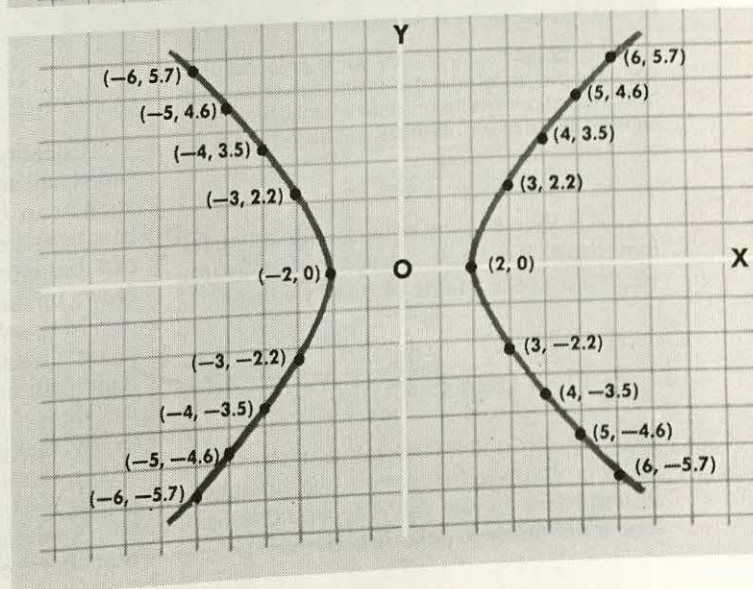
10. This figure is an ellipse. Its equation is $4x^2 + 9y^2 = 36$.

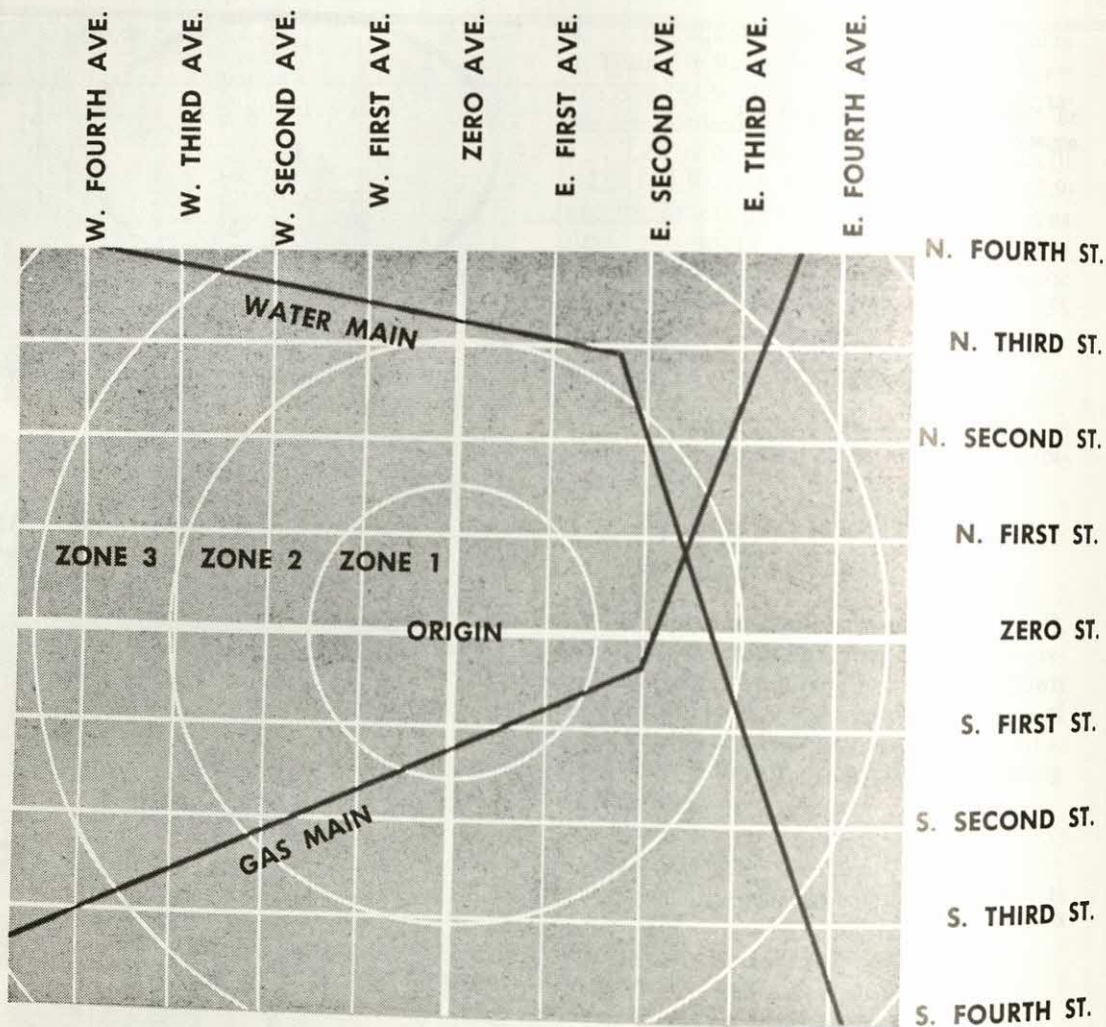


11. A parabola with the equation $x = y^2 - 4$. It is an open curve.



12. A hyperbola ($x^2 - y^2 = 4$). Its two branches are open curves.





13. Analytic geometry can be put to practical use. For example, water and gas mains can be located by the equations of straight lines with the town center used as the origin, as in the illustration above.

$x^2 - y^2 = 4$ is the equation for a typical hyperbola. To draw a graph of the figure, we first prepare a table of x and y values:

| | | | | | |
|-----|---------|-----------|-----------|-----------|-----------|
| x | ± 2 | ± 3 | ± 4 | ± 5 | ± 6 |
| y | 0 | ± 2.2 | ± 3.5 | ± 4.6 | ± 5.7 |

Then we plot the values given in the table, and we obtain a curve with two branches, as shown in Figure 12. The branches extend indefinitely in both directions.

USING THE EQUATIONS

Equations of geometric figures can be added and subtracted and many other operations can be performed with them. The equations resulting from such operations can be interpreted by means of graphs drawn up on the basis of x values and y values. In this way, algebra can be used to discover geometric relationships. As the American mathematician Eric T. Bell put it: "Henceforth algebra and analysis [that is to say, analytic geometry] are to be our pilots in the uncharted seas of space and its geometry."

Analytic geometry has also been put to practical use. For example, in many towns

gas and water mains can be located by the equations of straight lines using the center of the town as origin (Figure 13). Concentric circles can be used to give zone distances from the center of the town. The intersections of circles and lines will describe the approximate location of breaks in water or gas mains.

SOLID ANALYTIC GEOMETRY

To locate a point in three-dimensional space, we must give the distance above or below the plane formed by the x -axis and y -axis. We add a third axis—the z -axis, which extends straight up and down (Figure 14). Suppose we wish to locate point A , shown in the figure, in a three-dimensional system. A is $2\frac{1}{2}$ kilometers east of the origin (x -distance), 2 kilometers south of the origin (y -distance), and $\frac{1}{2}$ kilometer up (z -distance). We would locate it by giving the x -, y -, and z -, co-ordinates (2.5, 2, .5). To locate a point in solid analytic geometry, therefore, we need three ordered numbers— x , y , and z —as our co-ordinates. In the three-dimensional world, if some point is taken as the origin, any other point in space can be represented by an ordered triplet of numbers and can be definitely located.

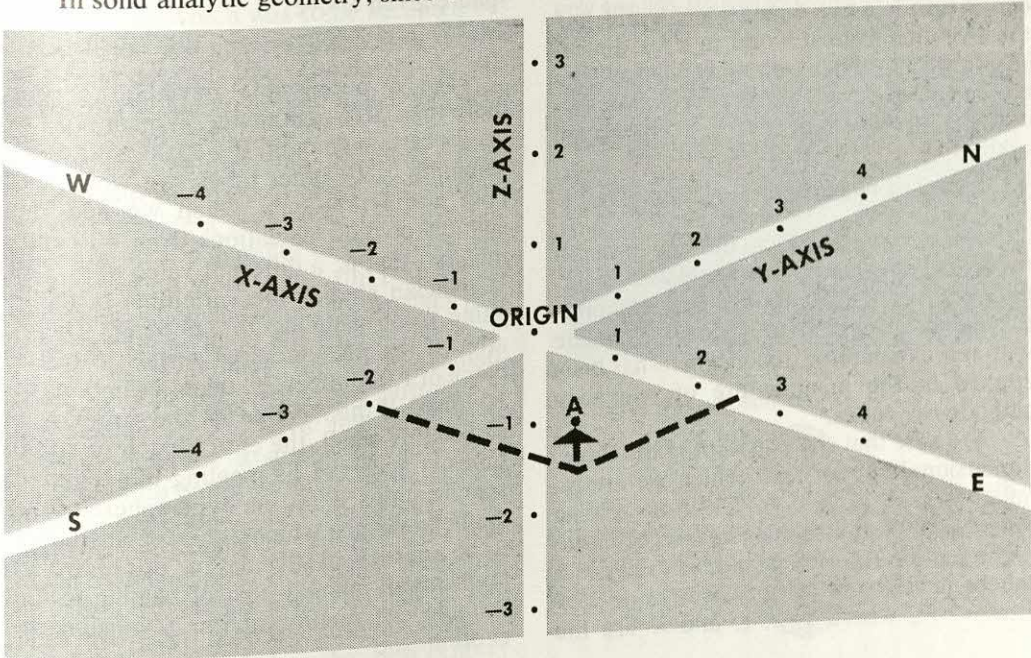
In solid analytic geometry, since there

are three dimensions, there will be three unknowns to be related in our equations. Thus the equation of a plane is $ax + by + cz = k$, where a , b , c , and k are given numbers. If the co-ordinates of a point are such that they satisfy the equation of a given plane, the point will be located on the plane. The co-ordinates of points that are located outside the plane will not satisfy the equation.

The equation of a sphere is $x^2 + y^2 + z^2 = r^2$, r being the radius. It is the extension of the circle equation of plane analytic geometry: $x^2 + y^2 = r^2$. The equation $x^2 + y^2 + z^2 = 25$ represents a sphere whose center is at the origin and whose radius is 5 units long. All the other common solids can also be represented by an equation or a series of equations in three unknowns— x , y , and z . Solid analytic geometry uses such equations to study the points, lines, surfaces, and solids that exist in space.

HIGHER GEOMETRY

The mathematician does not confine himself to the analysis of the three-dimensional world, in which an ordered triplet of 14. In solid analytic geometry, three axes— x , y , and z —correspond to the three dimensions. Point A in the diagram below, is $2\frac{1}{2}$ kilometers east of the origin, 2 kilometers south, and $\frac{1}{2}$ kilometer up.



numbers represents a point. The mathematician asks the question: "What would an ordered quadruplet of numbers represent?" Then he answers the question by saying that (3, 4, 2, 5), say, would locate a point in four-dimensional space. The equation $x + y + z + w = 3$ would represent the equation of what our mathematician would call a hyperplane; (3, 4, 2, 5) would be a point on this plane. What sort of thing would this hyperplane be? The mathematician does not know, because he has never seen one. That, however, does not prevent him from working out and using the equation of such a plane and from discovering the properties of the plane by various algebraic operations.

How can one maintain that a four-dimensional world does not exist? Suppose we were two-dimensional creatures living in a two-dimensional world. We would be quite unaware of a three-dimensional world, of which our own limited world would form a part. Yet this three-dimensional world does exist, as we all know. Is it not entirely possible that creatures like ourselves, confined as far as we know to a world of three dimensions, are really living in a four-dimensional world, which our senses cannot perceive? It may be, too, that this four-dimensional world is embedded in turn in a five-dimensional world, and the five-dimensional world in a six-dimensional world, and so on ad infinitum. The mathematician finds such speculation fascinating, whether or not it is based on reality.

CERTAIN PRACTICAL APPLICATIONS

As a matter of fact, the equations of these higher-dimensional worlds have certain practical applications. For example, the equation of a hypersphere—a sphere with more than three dimensions—has been applied in the manufacture of television cathode-ray tubes.

We saw that the equation of the three-dimensional sphere with which we are familiar is $x^2 + y^2 + z^2 = r^2$, r^2 representing the square of the radius. By analogy with this equation, the mathematician defines a sphere in four dimensions as $x^2 + y^2 + z^2 + w^2 = r^2$, where x , y , z , and w are the

four dimensions. This is the first hypersphere, and we make no effort to picture it. We then continue with the equations of a five-dimensional sphere, a six-dimensional sphere, and so on. We can define a hypersphere of n dimensions as

$$x_1^2 + x_2^2 + x_3^2 + \dots + x_n^2 = r^2$$

where x_1 , x_2 , x_3 , and so on up through x_n represent the different dimensions.

Now let us see how this equation can be applied to the manufacture of television tubes. The manufacturer of such tubes would like to have them all of exactly the same length—say 40 centimeters. Since, however, no tubes are ever of exactly the same length, he will allow a deviation of not more than .012 centimeter. If the deviation is greater than this, the tube will not be satisfactory.

A sampling of 10 tubes is selected for testing purposes from a batch of 1,000. The length of each one of these tubes is measured, and the deviation from 40 centimeters noted. The deviation of each tube is to constitute a dimension; hence we have 10 dimensions in all. We square each of the deviations and then add them together, in accordance with the equation for a hypersphere given above:

$$x_1^2 + x_2^2 + x_3^2 + \dots + x_n^2 = r^2$$

For a permissible deviation of not more than .012 centimeter, r^2 must not exceed a certain amount. If we take a different sampling of 10 other tubes, we would expect a different sum when we add the squares of their deviations from 40 centimeters—that is, a different value of r^2 . The statistician studies the variations of the r^2 value from sampling to sampling. Thereafter, if the manufacturer wishes to test a batch of 1,000 tubes, he takes a sampling of 10 tubes. If the r^2 value for the sampling is sufficiently small, the entire batch of 1,000 tubes is accepted as meeting the required specifications. Thus the hypersphere of the mathematicians, which at first seemed only to be impractical speculation, does have an application—namely that of helping to determine the quality of our television reception.

NON-EUCLIDEAN GEOMETRY

by Howard F. Fehr

The geometry presented in the articles Plane Geometry and Solid Geometry and that which is commonly taught in secondary schools is called Euclidean geometry. It is so named because it is based on the system established by the Greek mathematician Euclid and taught by him in Alexandria, Egypt, about 300 B.C. He established a logical series of theorems, or statements, which were so arranged that each one depended for its proof on (1) the theorems that preceded it and on (2) certain assumptions, or postulates. Euclid called these assumptions common notions and he accepted all of them without proof.

In order to prove some of the theorems in the first part of his system, Euclid found it necessary to assume that through a point outside a given line, only one line could be drawn parallel to the given line (Figure 1). He tried hard to prove that this was so, but failed. Finally he had to consider the statement about parallel lines (or rather, an equivalent statement) as a common notion, or postulate, which was to be accepted because it was self-evident. The entire system of Euclid depends upon the validity of this particular common notion, which has been called *Euclid's postulate*.

In the centuries that followed, many mathematicians tried to prove that through a point outside a given line, only one line could be drawn parallel to the given line. They were no more successful than Euclid had been. They had to accept Euclid's postulate as self-evident but not proven.

In modern times, not all mathematicians have conceded that the postulate is self-evident. They have felt justified in making different assumptions, and they have built up entire geometries based upon these assumptions. The 19th century German mathematician G. F. Bernhard Riemann based his geometry, called Riemannian geometry, on the postulate that no two lines are ever parallel. Another type of geometry was created in the 19th century

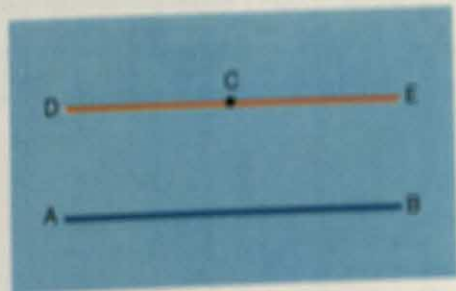
by the Russian Nikolai Ivanovich Lobachevsky and, independently, by the Hungarian Janos Bolyai. It was based on the assumption that at least two lines can be drawn through a given point parallel to a given line. The geometries of Riemann and of Lobachevsky-Bolyai are known, therefore, as *non-Euclidean*.

DIFFERENT ASSUMPTIONS

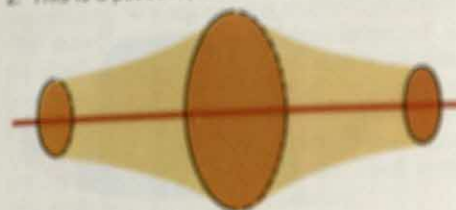
To understand the difference between Euclidean and non-Euclidean geometries, let us consider the shortest distance between two points in (1) a plane, or flat surface, (2) a sphere, and (3) a pseudosphere, which is shown in Figure 2 and which suggests two wastepaper baskets joined together at their tops. The shortest distance between two points on any kind of surface is called a *geodesic*.

On a plane, a hemisphere and a hemipseudosphere (half of a pseudosphere), we

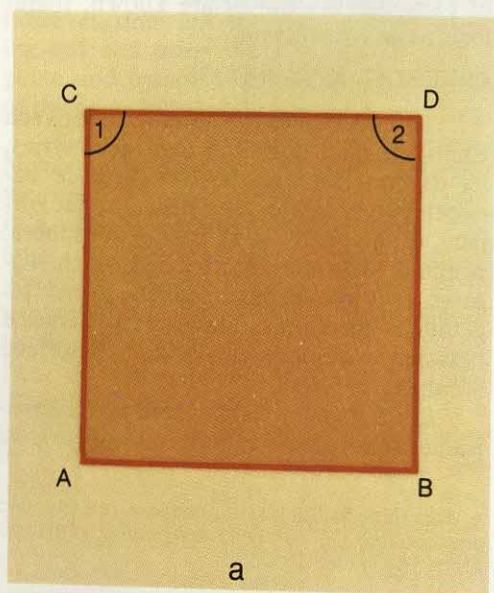
1. According to Euclidean geometry, one line and only one line, DE, can be drawn parallel to the line AB through point C.



2. This is a pseudosphere.



measure the same distance AB , using a geodesic (the shortest possible distance) in each case (Figure 3). On a plane, such as that shown in Figure 3a, the geodesic is a straight line. On a hemisphere (Figure 3b), it is an arc forming part of a great circle (the largest circle that can be drawn on a sphere). On the hemipseudosphere (Figure 3c), AB is on the circle that marks the "waist" of the pseudosphere.



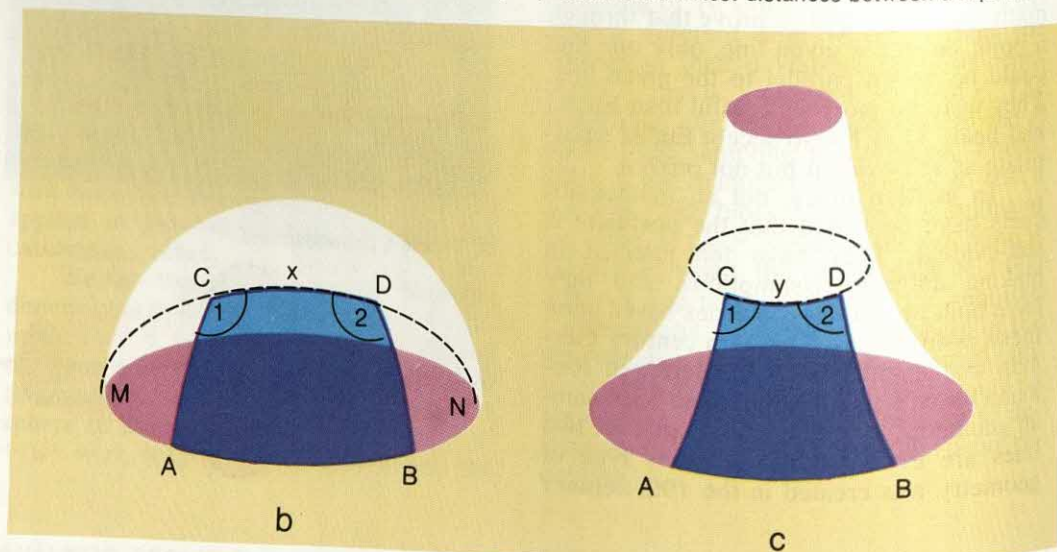
On the plane, hemisphere and hemipseudosphere, we draw geodesic AC , perpendicular to AB at A , and geodesic BD , equal to AC in length and perpendicular to AB at B . Finally, we draw the geodesic CD , connecting C and D . On the plane, CD is a straight line; on the hemisphere, it is an arc, CxD , forming part of the great circle $MCDN$; on the hemipseudosphere, it is the arc CyD .

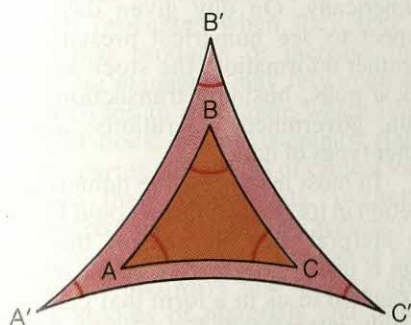
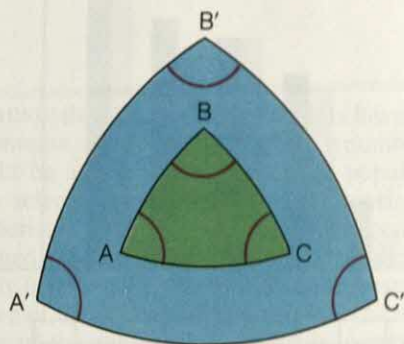
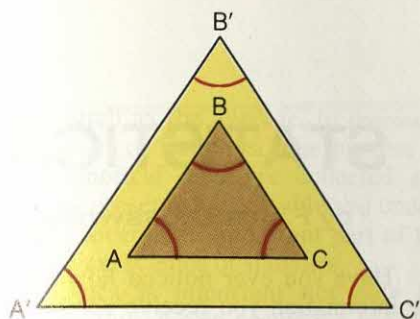
How large are the angles 1 and 2 in Figures 3a, 3b, and 3c? The answers to these questions will provide an insight into the differences between Euclidean geometry and the non-Euclidean geometries of Riemann and Lobachevsky-Bolyai.

In the case of the plane in Figure 3a, it would seem to be obvious that angles 1 and 2 are right angles (90° angles). But we cannot prove this as a geometric theorem unless we first agree that only one line passing through C —that is, the line CD —is parallel to line AB . If we make this assumption, we are accepting Euclid's postulate.

On the hemisphere (Figure 3b), angles 1 and 2 are apparently obtuse angles—that is, greater than 90° . Can we prove this? Not unless we first assume that every geodesic through C will meet line AB at two points. If we accept this, we can prove that angles 1 and 2 are greater than 90° .

3. In a, $ABCD$ has been drawn on a plane. In b, it has been drawn on a hemisphere; in c, on a hemipseudosphere (half of a pseudosphere). In each case, AC and BD are equal and they are perpendicular to line AB . All the lines shown here are geodesic; that is, they represent the shortest distances between the points.





6. Above: In the geometry of Lobachevsky-Bolyai, the sum of the angles of a triangle is always less than 180° . It decreases as the area increases.

Riemann assumed that even in a plane, any line (such as CD in Figure 3a) drawn through an external point (such as C) will meet any other line (AB) at two points. Hence there are no parallel lines in his geometry. He developed a perfectly logical set of theorems based on this assumption.

"But," you will say, "anyone can see that line CD in Figure 3a is parallel to line AB and that the two lines will never meet." But as long as you cannot prove that this is so, you cannot deny that it is perfectly logical to develop a new kind of geometry based on another assumption.

Let us now examine angles 1 and 2 in the hemipseudosphere (Figure 3c). These angles are apparently acute—that is, less than 90° . Again, we cannot prove that this is so unless we make an assumption—in this case, that there are two geodesics through C that never meet AB . Lobachevsky and Bolyai made this assumption and applied it to all kinds of surfaces, including those of spheres and planes. Their geometry de-

4. Above left: In the geometry of Euclid, the sum of the angles in a triangle, such as ABC , is 180° . If the area of such a triangle is increased so that we have the triangle $A'B'C'$, the sum of the angles is still equal to 180° .

5. Above right: According to the geometry of Riemann, the sum of the angles of a triangle, such as ABC , is always greater than 180° . It increases as the area of the triangle increases, as in the large triangle $A'B'C'$.

scribes a world quite different from that of Euclid and Riemann—a world in which through every external point there are two lines parallel to a given line.

Of course, three geometries based on such different assumptions are bound to show many striking points of difference. Consider, for example, the sum of the angles of any triangle. In the Euclidean geometry, the sum is 180° . It always remains the same no matter how much the size of a given triangle increases (Figure 4). In the Riemannian geometry, the sum of the angles of a triangle is always greater than 180° . As the area of the triangle increases, the sum of the angles increases (Figure 5). In the geometry of Lobachevsky-Bolyai, the sum of the angles of a triangle is always less than 180° and decreases as the area of the triangle increases (Figure 6).

One cannot say that one of these geometries is correct and that the others are incorrect. Rather, they are different explanations of space, based on different assumptions. We accept the one which offers the most satisfactory interpretation of a particular phenomenon with which we are concerned. For example, the Riemannian geometry has given a better explanation of certain astronomical aspects of Einstein's relativity theory than either the Euclidean or Lobachevsky-Bolyai geometry.



STATISTICS

by F. Joe Crosswhite

Have you ever noticed how much of the information you receive comes to you through numbers? Particularly since the coming of age of electronic computers, more and more kinds of information are being coded, processed, and presented numerically. On any given day, we can expect to see numerical presentations of weather information, the stock market, political polls, business transactions, census data, government operations, and many other types of data.

In most instances, the numerical information in its original form would be difficult to interpret. For this reason, the information is usually organized, summarized, and presented to us in a form that can be more readily interpreted. Frequently this is accomplished by reducing the numerical data to a table or graph or by reporting one

number, such as the average, to represent an entire set of numbers. The process by which numerical data are collected and eventually presented in a usable and understandable form is an important part of the mathematical science of statistics.

Statistics is important not only for communication; it also provides a basis for decision-making. The government makes extensive use of statistics in estimating its budget needs and setting its tax rates. Statistics enables manufacturers to compare production processes when they seek to improve their products or increase their profits. Store managers may rely upon statistical analyses to determine which items they should stock. Scientists employ statistics in comparing the effects of critical variables upon their experiments. Insurance companies rise and fall on the accuracy of their statistical predictions. Engineers base the design of highways and bridges upon statistical studies of materials and traffic. School officials may modify their curricula on the basis of statistical analyses of student achievements and needs. The list of such decision-making uses of statistics is almost endless.

COLLECTION OF DATA

The science of statistics involves a variety of tasks. Even the seemingly simple business of collecting numerical data requires careful study. Obviously the conclusions of a statistical study can be no more reliable than the figures upon which they are based. The statistician must be sure that the data collected are accurate, relevant to the problem being studied, and representative of the problem. Invalid conclusions drawn from statistical evidence are often due to inadequacies in the data collected. Although we cannot go into the matter of data collection in detail, its importance is so great that we should be at least aware of some of the problems involved.

First, the *population* to be studied must be well-defined. What do we mean by "population" here? To the statistician, it may consist of a set of cities, or automobiles, or books, or even scientific experiments. In fact, the population for a statisti-

cal study might be any set of objects having a common characteristic to which a number might be assigned. Of course, the population selected must supply the appropriate numerical data for the problem being studied. If the population of a statistical study is not well-defined or is not representative of the problem being studied, the results of the study will be difficult to interpret or apply. For example, surveys of the voting preferences of high-school students would be of questionable value in predicting the results of an election, since few high-school students can vote.

Once the population has been identified, the particular characteristics to be studied must be represented numerically. Sometimes the numerical data are already available in recorded form. For example, if you wanted to study the rainfall in your city over the past year, you could probably obtain the needed data from your local weather bureau. In this case, the population might be defined as the set of days in the year.

Sometimes the numerical data may be obtained by a simple counting process. You might be interested, for example, in a study of the books in a school library. This project would involve counting the number of volumes devoted to each of several subjects.

More often, the data for a statistical study are obtained by measuring some common characteristic of the population being studied. If the population were a set of scientific experiments, the scientist might be concerned with such characteristics as time, temperature, volume, and mass. In each instance, he would need to use a suitable measuring instrument to assign a number to the characteristic.

In some cases, no measuring instrument is available and the investigator must create a measuring device. For this purpose, the investigator may construct an examination. This then becomes the instrument that enables him to assign a number to measure student achievement. As with any other measuring instrument, accuracy is a prime consideration. The investigator would need to know how well the number

TABLE I

Thirty-Five
Scores on
a Spelling
Test

| | | | | | | |
|----|----|----|----|----|----|----|
| 16 | 18 | 16 | 14 | 12 | 13 | 18 |
| 19 | 12 | 17 | 15 | 16 | 14 | 16 |
| 14 | 16 | 20 | 16 | 17 | 15 | 15 |
| 16 | 15 | 14 | 18 | 15 | 16 | 13 |
| 17 | 13 | 16 | 17 | 15 | 19 | 17 |

assigned represents the true value of the characteristic being measured—in this case, student achievement.

The ultimate value of a statistical study depends to a large extent upon the quality of the measuring instrument that is employed. For this reason, the construction and evaluation of such an instrument is often a critical task in a statistical study.

Sometimes it is possible to obtain numerical data about each member of a population that is being studied. When this is true, the data are completely representative of the population, and the task of the statistician is to describe the numerical data obtained. This branch of statistics is called *descriptive statistics*.

SAMPLING A POPULATION

Often it is necessary or practical to collect data only for a *sample* of the population and to make *statistical inferences* about the population itself. An inference is a conclusion about the unknown based upon something that is known. A statistical inference, or course, is one based upon statistical data. When data are available only for a sample, the sample represents the known and the population the unknown. Any subset of a population would constitute a sample, but statistical inferences are valid only when the sample is representative of the population. Many techniques are employed

by statisticians to insure that the samples they select are representative.

When each member of the population has an equal chance of being chosen, we have what is called a *random sample*. This is usually assumed to be representative of the population. In some special problems, the statistician uses a *stratified sample*, which insures that specific segments of the study population are represented in the sample. The process of identifying a representative sample is a critical task in many statistical studies. In the examples that follow, we will assume that the samples used are representative of the populations from which they have been selected.

Let me summarize the above remarks about collecting data by considering the following example. Suppose that the population being studied is the set of students in a given grade. If each of these students was assigned to an English class by a random process, the English class would constitute a representative sample of the population. If you were to measure the height, weight, or age of each student in the class, or record his score on a particular test, or count the number of people in his family, you would obtain a set of numbers. These numbers could then become the data for a statistical study. The data could be used to describe the English class (the sample). They could also be used to make *estimates* or *infer-*

ences about the total set of students in the grade (the population).

ORGANIZING THE DATA

Once data have been collected, they must be arranged in some systematic order before a useful interpretation can be made or conclusions drawn. Sometimes a simple table or graph can be quite helpful as a first step toward the statistical analysis of numerical data.

The numerical data presented in Table I are scores obtained by thirty-five students in a spelling test. Each score represents the number of words that have been spelled correctly by a given student. In this case, the word "score" is used in its usual sense. However, regardless of the nature of the numerical data, statisticians often use the term *raw score* to indicate the individual numbers obtained as a basis for a statistical study.

It is difficult to make any useful interpretation of the data in Table I. The simplest way of organizing these data would be to arrange the scores in numerical order. It is common to record only the different raw scores (in this case, 16, 18, 14, and so on) and to note the *frequency* with which each score occurs. Table II is a frequency table presenting the same data that appear in Table I.

Even a cursory examination of Table II permits some elementary interpretation

of the data. We can easily observe the highest and lowest scores (20 and 12) and the most frequent score (16). We can even begin to have some feeling for the way the scores seem to cluster about a central point—in this case, the score 16.

Further clarification of the data may be obtained by translating Table II into a graphical form. Figure 1 is a common type of graph used to present frequency distributions. The numbers below the horizontal line represent scores; the numbers at the left represent the frequency distribution—that is, the number of times each score occurs. This is a frequency graph.

The frequency polygon shown in Figure 2 is based on the same idea. In this case, we may consider that lines have been drawn perpendicular to the scores on the bottom line and to the frequencies at the left. A point indicates the intersection of a score line and a frequency line. Thus we have a point where the score 12 and the frequency number 2 meet, and a point where the score 16 and the frequency number 9 meet. The points are connected by straight lines. A line is drawn in each case between the point representing the lowest score and the score immediately preceding on the horizontal line; also between the point representing the highest score and the score immediately following on the horizontal line.

An examination of these two graphs

TABLE II
Frequency Distribution of Spelling Scores

| Score | Frequency |
|--------|-----------|
| 20 | 1 |
| 19 | 2 |
| 18 | 3 |
| 17 | 5 |
| 16 | 9 |
| 15 | 6 |
| 14 | 4 |
| 13 | 3 |
| 12 | 2 |
| n = 35 | |

Note: "n" stands for the number of scores.

TABLE III
Grouped Frequency Table of 200 Spelling Scores

| Interval | Frequency |
|----------|-----------|
| 23-25 | 8 |
| 20-22 | 19 |
| 17-19 | 44 |
| 14-16 | 60 |
| 11-13 | 40 |
| 8-10 | 21 |
| 5-7 | 8 |
| n = 200 | |

will reveal exactly the same information available in Table II. For some people, the graphical form is easier to interpret than the tabular form. In particular, the tendency of scores to cluster around a central point becomes more apparent when the data are presented pictorially.

GROUPING RAW SCORES

In many statistical studies, the number of different raw scores obtained is so large that it becomes cumbersome to work with all the different scores. In these cases, it is common to condense the data by grouping the raw scores into *class intervals*. Instead of considering each score individually, we combine a certain number of adjacent scores to form an interval. Thus we would combine the individual scores 23, 24, and 25 to form the interval 23–25. Intervals are treated in much the same way as we would treat individual raw scores.

A grouped frequency table, based on intervals, is shown in Table III. It presents scores made by 200 students in the same spelling test for which thirty-five scores were reported in Table I. We could think of the data in Table I as representing the scores of one of six randomly assigned English classes whose total scores are reported in Table III.

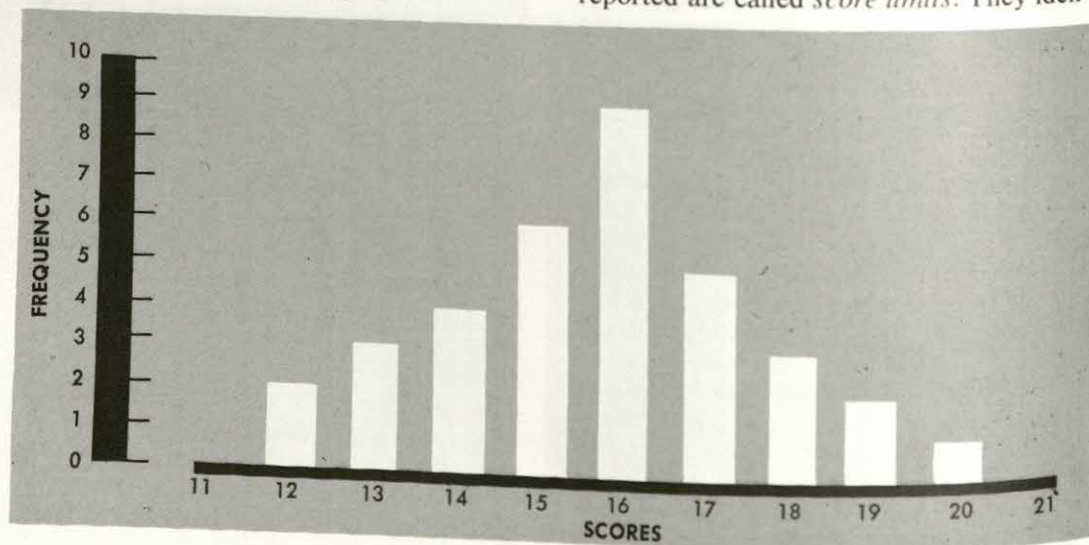
When the number of possible raw

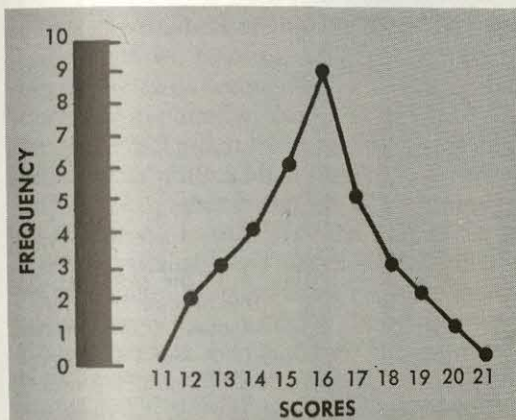
scores is large, the grouping of data into appropriate intervals enables the investigator to work with a manageable number. The seven intervals in Table III may reflect as many as three times that number of different raw scores. At the same time, this grouping method has the disadvantage of obscuring individual scores. Thus only one score may be represented in the interval 23–25, while all three scores might be represented in another interval. We have to ignore this consideration when working with class intervals. Once the data have been compressed by grouping, in all subsequent analysis and computation, we must treat individual scores as if they were evenly distributed throughout the interval to which they belong.

Graphical representations of grouped frequency tables are very similar to those presented earlier for ungrouped data. One special kind of graph, the *frequency histogram*, is worthy of note. To construct a histogram from Table III, as in Figure 3, the frequency of scores in each interval is represented by a rectangle with its center at the mid-point of the interval, its height equal to the frequency, and its width equal to the width of the interval—with a difference.

Here it is necessary to distinguish between *score limits* and *real limits*. When an interval is identified as 14–16, the limits reported are called *score limits*. They iden-

1. Frequency graph of data in Table II.





2. Graphs are often easier to interpret than tables.

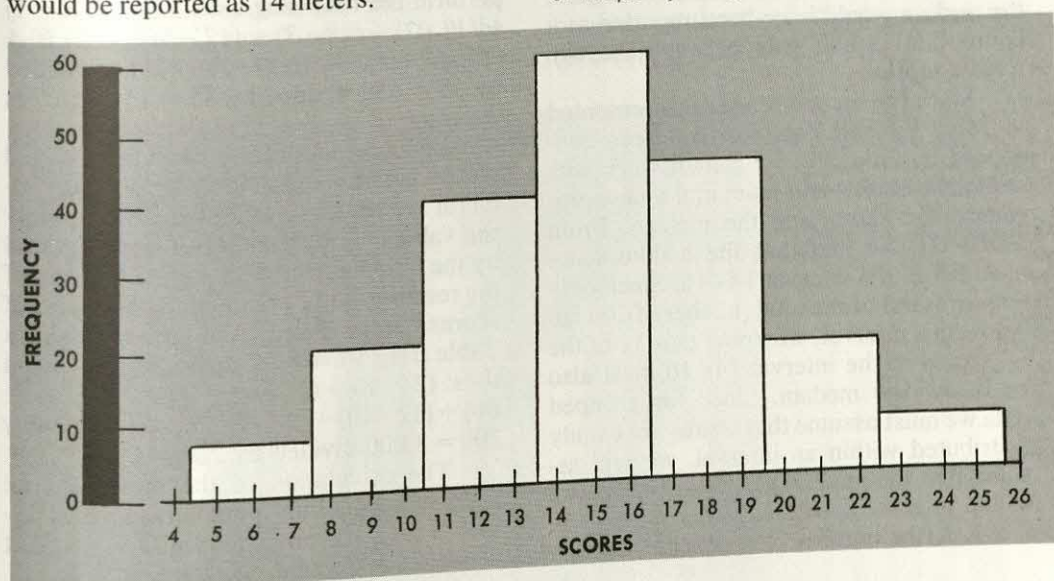
tify the lowest score and highest score that belong to the same interval. For purposes of mathematical treatment and graphical representation, it is common to use the real limits 13.5–16.5 to identify this same interval. The interval is represented as extending halfway to the scores immediately preceding and following. Such an interpretation is consistent with the way we usually report measurements. For example, if we were measuring to the nearer meter, any measurement between 13 meters, 50 centimeters and 14 meters, 50 centimeters would be reported as 14 meters.

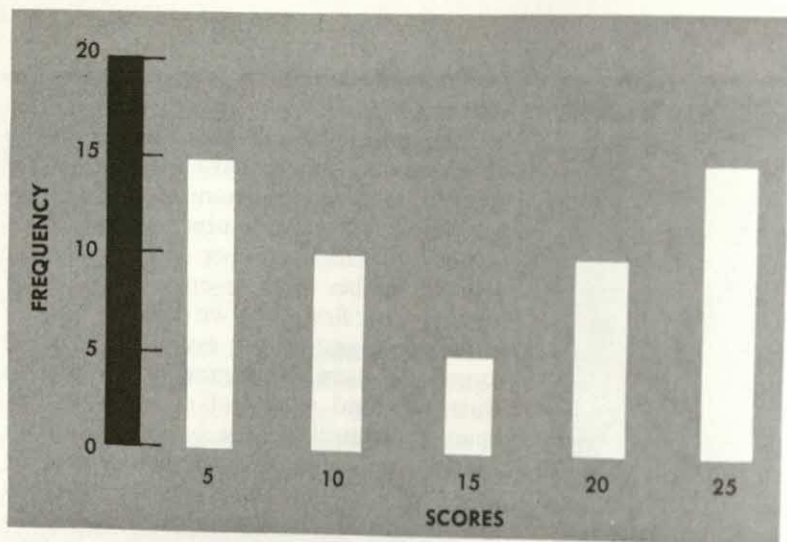
ANALYZING THE DATA

Tables and graphs can help us obtain considerable understanding of a set of scores. However, for many purposes, it is more desirable to try to represent the set of scores by a single number. When selecting a single number to represent a whole set of numbers, the first thing we usually think of is the average. As we noted earlier, the scores we have been examining seem to cluster around a central point. It is this point of central tendency which statisticians identify when they report an average score. In statistics, there are several types of averages. Three are common in statistical analysis—the mode, the median, and the mean. Each is called a measure of central tendency. For the data with which we have been working in this article, the mode, median, and mean are close together. This is not always the case. They may be appreciably different.

The *mode* is quite easily identified from a frequency table or frequency graph. It is the score that occurs most frequently—in a sense, the most popular score. In the ungrouped data presented above, we have already noted that the most frequent score is 16. This can be determined from Table II, Figure 1, or Figure 2. Thus 16 is the mode

3. Frequency histogram of data in Table III.





4. In the graph of this distribution there are 2 scores—5 and 25—that have the highest frequency. Compare it with the graph in Figure 5.

of this set of scores. In the grouped data, the *crude mode* would be identified as the midpoint of the interval with the highest frequency. From either Table III or Figure 3, we can see that the interval of highest frequency is 14–16. Hence 15 would be the crude mode of this distribution since it is the midpoint of the interval. Since individual scores have been obscured, we cannot be sure that it is actually the most frequent individual score, but it is our best estimate of the most frequent score.

The *median* is the middle score in a set of scores. If we examine the set of scores in Table II, we see that the median is 16. If there had been an even number of scores, the median would have been reported as a figure that is half way between the two middle scores.

Since the second set of data, presented in Table III and Figure 3, has been condensed by grouping, we cannot work with individual scores and must find a new procedure for identifying the median. From Table III, we find that the middle score must fall in the interval 14–16. Since only 69 scores out of the total number of 200 fall below this interval, we know that 31 of the 60 scores in the interval 14–16 must also be below the median. Since for grouped data we must assume that scores are evenly distributed within an interval, we will assume that the median lies $\frac{31}{60}$ of the width of the interval above its lowest boundary. $\frac{31}{60} \times 3$ (the number of scores in the inter-

val) = $\frac{93}{60} = 1.5 +$. We add this number to the left end point of the interval 14–16. We noted previously that this point is 13.5 since an interval begins and ends halfway between two scores. Hence we have $13.5 + 1.5 = 15$. The median, then, is 15.

The *mean* is the most commonly used measure of central tendency, and it the average most of us think of first. It is found by dividing the sum of all the individual scores by the number of scores in the set. This calculation can be shortened when a frequency table is available if we multiply each score by its frequency and then find the sum. From Table II, for example, we would perform the following calculation: $(20 \cdot 1) + (19 \cdot 2) + (18 \cdot 3) + (17 \cdot 5) + (16 \cdot 9) + (15 \cdot 6) + (14 \cdot 4) + (13 \cdot 3) + (12 \cdot 2)$ divided by 35 = 550 divided by 35 = 15.7. 15.7 is the mean.

When computing the mean for grouped data, we assume that the scores in any interval are evenly distributed. We multiply the value of the midpoint of each interval by the frequency, and we divide the sum of the resulting number by the total number of scores. In the case of the grouped data in Table III, we could compute the mean thus: $(24 \cdot 8) + (21 \cdot 19) + (18 \cdot 44) + (15 \cdot 60) + (12 \cdot 40) + (9 \cdot 21) + (6 \cdot 8)$ divided by 200 = 3,000 divided by 200 = 15.

The calculation of the mean can be simplified by using more advanced methods. These are mathematically equivalent to the above calculations.

The mode is used less frequently than either the median or the mean. It is useful only when we want to identify the number occurring most frequently in a set of numbers. As a matter of fact, if the mode is to be truly meaningful, one number in a set must occur quite a bit more frequently than any other number in the set. The advantage of the mode is that, like the median, it is easy to identify and understand. But the term "mode" is sometimes ambiguous because there may be more than one score with the "highest frequency" (see Figure 4). Also the mode is not too reliable as an indication of central tendency because the most popular score is not always near the center of a given distribution.

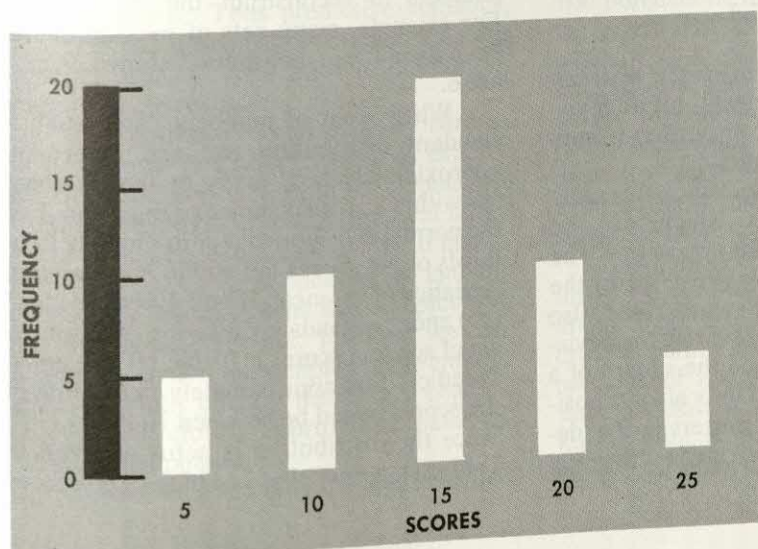
The chief advantage of the median is that it is not affected by extreme scores. The "average" income in a community, for example, is often more accurately reflected by the median than the mean because the value of the median is not influenced by a few very high or very low incomes. The idea of the median is closely related to the concept of percentiles—a type of "score" students receive on certain standard tests in school. The median corresponds to the 50th percentile. Other percentiles can also be used in connection with tests. They are valuable as a basis for comparing individual scores with other scores in a distribution.

For most purposes, the mean is the best measure of central tendency. It is the only one of the three measures that depends

upon the numerical value of each score in a distribution. It is a reliable indicator of "central tendency" because it always identifies the "balancing point" or "center of gravity" in the distribution. Since the mean lends itself better to mathematical computation, it is more suitable for deriving other statistical measures. For example, the means of two sets of data can be used to compute a mean for the combined set of data. This cannot be done with the mode or the median.

However, the mean can give us information only about the central point in a distribution. To understand a set of scores more fully, we also need to know how the scores spread out around this central point. For this reason, statisticians develop measures of *dispersion* or *variability*.

The simplest measure of distribution is the *range*, which is defined as the difference between the highest and lowest scores in a distribution. The range of spelling scores reported in Table II is easily identified. We simply subtract the lowest score (12) from the highest score (20). Since the range is sensitive only to the two extreme scores in a distribution, it is not considered a very satisfactory measure of dispersion. Its weakness is dramatized in the two distributions whose graphs are presented in Figures 4 and 5. Both the mean (15) and the range (20) are identical for these two distributions; yet the distributions are obviously different.



5. In the distribution graph here, the mean (15) and the range (20) are identical with those in Figure 4. But the distributions are different.

As with the median and the mode, the weakness of the range is that it does not take into account the numerical value of each score. A natural measure of dispersion involving every score is the average difference between the individual scores and their mean. As we have seen, the mean serves as a "balancing point" for a distribution. Hence we can measure the deviations from the mean in terms of positive and negative differences. Deviations from the mean in one direction would be positive; in the other, negative.

Since the mean is the "balancing point" or "center of gravity" for a distribution, the sum of the deviations from the mean is 0. (This is because some of the differences are positive and others are negative.) That is why we must employ the mathematical notion of *absolute value* to treat the deviations from the mean as distances without regard to direction—that is, without indicating whether they are positive or negative. The average of the absolute values of the differences between individual scores and the mean is called the *mean deviation* and is a simple and accurate measure of dispersion. For the distribution pictured in Figure 4, the calculation of the mean deviation would appear as follows: $15 \cdot |5 - 15| + 10 \cdot |10 - 15| + 5 \cdot |15 - 15| + 10 \cdot |20 - 15| + 15 \cdot |25 - 15|$ divided by 55 (frequency). It would be equal to $15 \cdot 10 + 10 \cdot 5 + 5 \cdot 0 + 10 \cdot 5 + 15 \cdot 10$ divided by 55. This would give us

$$\frac{400}{55} \approx 7.3$$

A similar calculation for the distribution in Figure 5 would yield a mean deviation of 4.0. Thus greater dispersion is indicated for the figure in Figure 4.

The use of absolute values presents mathematical difficulties which can be avoided by using another measure. The positive and negative signs that led to the introduction of absolute values could also have been eliminated by squaring the deviations from the mean, since the square of a positive or negative number is always positive. Such a procedure preserves the descriptive qualities of the mean deviation

while providing a measure that is easier to handle mathematically. Hence statisticians prefer to use the *standard deviation*, indicated by the symbol " σ ", as a measure of dispersion. The standard deviation is defined as the square root of the average squared deviation from the mean. Thus to calculate the standard deviation, we first find the average of the squares of the deviations. This number is called the *variance*. Suppose we wished to find the variance and standard deviation (σ) for the data in Table II. We would obtain the variance thus: $(20 - 15.7)^2 + 2(19 - 15.7)^2 + \dots + 3(13 - 15.7)^2 + 2(12 - 15.7)^2$ divided by 35 (total number of scores.) This would be equal to 128.15 divided by 35 \approx 3.66. The variance then is 3.66. We would then have

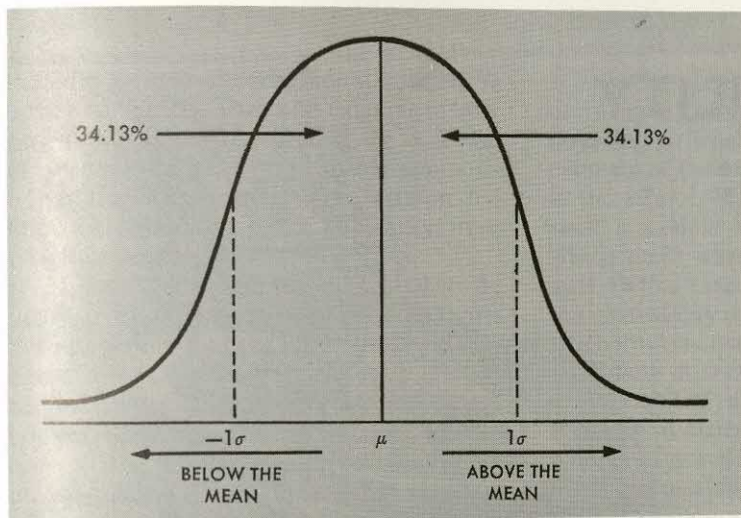
$$\sigma = \sqrt{\text{variance}} = \sqrt{3.66} \approx 1.9$$

Similar calculations would yield a standard deviation of 4.2 for the data in Table III. If we compare these two measures of dispersion, we see that the scores in Table II (the sample) are not so "spread out" as the scores in Table III (the population).

INTERPRETING THE DATA

Together, the mean and the standard deviation give us a reasonably clear picture of a distribution, because they describe both its central tendency and its dispersion. Sometimes, if we know the general nature of the distribution, we need only these two numbers to reconstruct the distribution. For example, many sets of measurements have the *normal* distribution shown in Figure 6.

When a set of numbers "fits" such a standard distribution, we can determine approximately how many of the numbers fall within a given distance of the mean. For the normal distribution, approximately two-thirds of the scores fall within one standard deviation of the mean. Thus, given the mean (15) and the standard deviation (4.2) for the set of spelling scores in Table III, we could "predict" that approximately two-thirds of the scores would be between 10.8 and 19.2. Since the distribution is actually given in Table III, we see that 144 of the 200 scores



6. A normal distribution curve. The symbol μ stands for the mean; σ stands for the standard deviation.

actually do fall in this interval. The "prediction" is fairly accurate since the number of scores is relatively large and they do "fit" the normal distribution.

The knowledge of such general models for distribution coupled with the laws of probability form the basis of *predictive statistics*. Both statistics and probability have to do with distributions and it is upon this common focus that we capitalize in predictive statistics. In probability, the sample space (population) is known and we predict the composition of a set of outcomes (sample). In statistical inference, the sample (set of outcomes) is known, and we infer the composition of the population (sample space.) Thus predictive statistics, also called statistical inference, can be thought of as an application of the laws of probability in reverse.

If we could be certain that the distribution of scores in a sample reflected exactly the distribution of scores in the population from which it was chosen, statistical inferences would be exact and simple to make. But even when a population is known, probability theory tells us that samples will not always be the same. The best we can hope for is that, if the sample is large enough and is carefully chosen, the sample characteristics will closely approximate those of the parent population.

Suppose we had only the data recorded

in Table II and wished to estimate the mean spelling score for all the English classes. Our best estimate would be the mean obtained for the sample: 15.7. But knowing that samples vary, we would hedge on this estimate. We would give a *confidence interval* within which we would expect the true mean of the population to fall. By assuming the total set of spelling scores to be normally distributed and applying basic laws of probability, we could determine that there is 95 per cent chance that the population mean falls in a given interval with its center at 15.7. The 95 per cent is a measure of the confidence or reliability we can place in our interval estimate. It means that 95 of every 100 populations from which a sample with the given characteristics might be chosen would have a mean within the determined interval. Because of the uncertainties involved in sampling, such an interval estimate, accompanied by a statement of the degree of confidence we can place in the estimate, is preferable to a single number approximation.

The process of establishing confidence intervals permits us to test hypotheses about a population. Modifications of this process permit us to make statistical comparisons of two samples drawn from the same population, to compare a sample to a known population, or to infer other characteristics of an unknown population.

PROBABILITY

by F. Joe Crosswhite



How often should we expect all three of the children in a family to be boys?

What are your chances of winning a sweepstakes contest or the door prize at a party?

How likely are you to land on "Boardwalk" in your next turn in a game of Monopoly?

If a batter averages 3 hits in 10 times at bat, what are his chances of getting 4 straight hits?

Have you ever wondered about the answers to such questions? Scientists often have occasion to study questions of this general type (though much more complicated than the four given above). In seeking the answers, they apply the mathematical theory of probability.

Probability is a mathematician's way of describing the likelihood that a certain event will take place. It is used to predict the outcome of an experiment when this outcome is governed by the laws of chance. Probability theory enables us to determine probable characteristics of a sample drawn

from a population whose characteristics are known. It also provides the basis for part of the related science of statistics. Statistical inference is an extension of probability theory. In it, the outcome of an experiment is used to estimate the conditions governing the experiment. We would be drawing a statistical inference if we were to make an educated guess about some population by examining the characteristics of a sample taken from that population.

Let us now consider some examples. Suppose we know that there are 400 boys and 200 girls in a certain school and we conduct the experiment of choosing one student from this school at random. In a random selection, each student would have exactly the same chance of being chosen. In this experiment, we would say that the probability of choosing a boy is $\frac{2}{3}$ and the probability of choosing a girl is $\frac{1}{3}$. The numbers $\frac{2}{3}$ and $\frac{1}{3}$ indicate that on the average we should expect to choose a boy two times out of three and a girl one time out of three.

If we repeated the above experiment fifteen times, our best estimate would be

that the sample chosen should contain $\frac{2}{3}$ boys and $\frac{1}{3}$ girls. This is an example of how probability is used to predict the outcome of an experiment. The probable composition of the sample was determined from the known composition of the population from which it was chosen.

Suppose that in a second school we do not know the proportion of boys and girls but know only that the total number of students is 600. To estimate the ratio of boys to girls, we might conduct an experiment consisting of selecting 20 students at random. If the sample chosen in this experiment contains 10 boys and 10 girls, our best estimate would be that there are equal number of boys and girls in the school—that there are 300 boys and 300 girls. In this case, the composition of the sample permits us to make a statistical inference about the composition of the population from which it was drawn.

As you see from these examples, probability and statistical inference are very closely related. The latter may be thought of as an application of probability in which the reasoning process is reversed. Statistical inference is only one aspect of the field of statistics, which also involves the collection, presentation, and analysis of data.

The examples given above illustrate the basic idea of probability and suggest how it may be applied, but they also may be misleading. They were oversimplified in order to avoid some of the more difficult problems which arise in the study of chance phenomena. In the first example, if we repeated the experiment a very large number of times, we might expect to choose a boy at random instead of a girl just about two times out of three. But the chances of selecting exactly 10 boys in a given sample of 15 are really quite small. This would happen only about two times in any ten trials. Most of the samples of 15 would be close to the theoretical ratio of $\frac{2}{3}$ (11 boys and 4 girls or 9 boys and 6 girls, for example), but some would not be close at all. We might even choose a random sample of 15 made up entirely of boys. In other words, the most probable outcome may not be very probable at all.

In the second example, the fact that our sample contained an equal number of boys and girls could lead us to a very poor estimate of the proportion of boys and girls in the school. We might even have drawn such a sample from the first school, in which there were twice as many boys as girls. In that case, we would have estimated the probability of choosing a boy as $\frac{1}{2}$ when it was actually $\frac{2}{3}$.

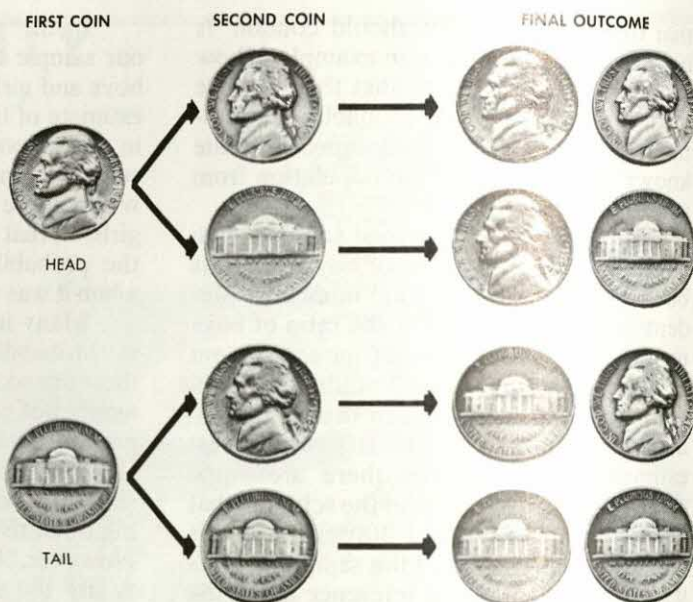
Many interesting and useful problems in probability are complicated because there are so many possible outcomes. If the number of possible outcomes is infinite, we must use calculus in our analysis. In this short introduction to probability theory, we shall deal with simple experiments producing a relatively small number of outcomes. However, the basic principles of probability are the same in simple problems as in more complicated ones. As you read, try to understand these principles, so that you can apply them to more complex situations. It will help you to grasp the subject if you work out all the situations given in the text as well as the special problems. Mathematics is not a spectator sport.

PROBABILITY AND GAMES OF CHANCE

Games of chance involving coins, cards, and dice provide us with simple experiments producing a small number of outcomes. The use of such experiments to illustrate the principles of probability is historically appropriate. Historians tell us that a seventeenth-century French gambler, the Chevalier de Méré, was interested in the odds involved in a game of chance played with dice. He decided to get in touch with a famous mathematician and scientist, Blaise Pascal, so that the latter might help him with his calculations. Pascal became intrigued with the interesting questions that arose in his study of De Méré's problem. He began a correspondence concerning this matter with other mathematicians and this led to the development of probability theory.

Consider the simple experiment of tossing coins. There are two possible outcomes when we toss a single coin—heads or tails. We would ignore any toss in which

1. Possible outcomes in the tossing of two coins. An analysis of the likelihood of each of the four possible outcomes appears in the text.



the coin came to rest on its edge. If the coin is in fair condition and if we toss it vigorously, it seems reasonable to say that heads and tails are equally likely outcomes. A mathematician would say that the probability of heads is $\frac{1}{2}$ —that is, that the coin would land heads one time in two on the average.

Suppose we complicate this experiment just a little by tossing the coin twice or, what comes to about the same thing, tossing two coins. We can see that three things might happen in this experiment. We could get two heads, or one head and one tail, or two tails. The diagram in Figure 1 illustrates the possible outcomes in this experiment.

Let H stand for “heads”, T for “tails.”

If we examine the diagram carefully, we see that there are really four individual outcomes, or elementary events, possible—HH, HT, TH, and TT—when we toss two coins. HT is not considered here as the same thing as TH. We might, for example, use a nickel and a dime as our two coins.

Since each of the four individual outcomes is equally likely to occur, we would expect to obtain HH one time in four. In other words, the probability is that we would obtain two heads one time out of four or, as a mathematician would indicate

it, $P(HH) = \frac{1}{4}$. Similarly, we would say $P(HT) = \frac{1}{4}$; $P(TH) = \frac{1}{4}$; $P(TT) = \frac{1}{4}$.

Because what happens to the first coin has no effect upon what happens to the second coin, we say that the two tosses are independent. When two events are independent, we can use their individual probabilities to compute the probability that both will happen in a single trial of an experiment. In this case, we could have used the probabilities associated with tossing a single coin—that is, $P(H) = \frac{1}{2}$; $P(T) = \frac{1}{2}$ —in order to compute the probabilities associated with tossing two coins. For example, $P(HT) = P(H) \cdot P(T) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$; $P(HH) = P(H) \cdot P(H) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$; the symbol \cdot stands for “times.” Since the event “one heads and one tails” could occur in two ways, either HT or TH, we would expect to obtain “one heads and one tails” two times in four on the average and would assign probability to this event.

We could also use the individual probabilities assigned to the outcomes HT and TH to compute the probability that either HT or TH will occur. When two events cannot both occur in a single trial of an experiment, the probability that one or the other will occur is the sum of their individual probabilities. Thus $P(HT \text{ or } TH) = P(HT) + P(TH) = \frac{1}{4} + \frac{1}{4}$

$= \frac{1}{2}$. In these examples, we obtain the probability of an event either by counting outcomes of an experiment or by using previously determined probabilities.

We can apply the same principles in other experiments. Consider the question "How often should we expect all three of the children in a family to be boys?" If we assumed that equal numbers of boys and girls are born (this is not quite true), we would say that the probability of a boy, $P(B)$, is $\frac{1}{2}$ and that the probability of a girl, $P(G)$, is also $\frac{1}{2}$. The possible outcomes for this experiment would be equal in number to the outcomes for tossing three coins—BBB, BBG, BGB, BGG, GBB, GBG, GGB and GGG. Since BBB occurs in only one of the eight possible outcomes, we would say $P(BBB) = \frac{1}{8}$. We could also have used the rule for computing the probability that all of several independent events will occur— $P(BBB) = P(B) \cdot P(B) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$. Either way, we would conclude that we should expect all three of the children in a family to be boys about one time in eight.

Some of the basic principles of probability were involved in the examples discussed above. Let us now examine these principles in more detail.

SAMPLE SPACES

To determine the probability of a particular outcome of an experiment, we must be able to identify all the possible outcomes. Mathematicians call the set of possible outcomes of an experiment a *sample space* for the experiment. In the simple examples that we shall consider, we shall usually find it convenient to list the sample spaces.

When we considered the experiment of tossing a single coin, our sample space consisted of the two possible outcomes H and T (heads and tails). When we extended the experiment to the tossing of two coins (or one coin twice), we used a sample space of HH, HT, TH, TT. In the experiment of choosing a single student from a known school population, the sample space involved was the set of all students in the school. Since we were concerned only with

whether we chose a boy or girl and not with which boy or girl, we could have used a sample space of only two elements—Boy and Girl. In this case, B could be used to stand for the set of boys and G for the set of girls. Each individual outcome of a sample space is also called a *point* or an *elementary event*.

Although a single experiment will produce only one set of individual outcomes, we may be able to use any one of several different sample spaces, depending on what we are investigating. For example, consider the experiment of drawing one card from a well-shuffled deck of cards. If we are concerned with the individual card drawn, we could use a sample space consisting of 52 elements—every single card in the deck. In the same experiment, we might be concerned only with the face value of the card shown. In that case, our sample space would consist only of the 13 elements (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King). Or we might be interested only in the suit drawn. In that case, we would use the sample space of Clubs, Diamonds, Hearts, Spades. If we were concerned only with color, we could use a sample space of only two elements (Red, Black). In most experiments, we use the sample space identifying the characteristic on which we wish to concentrate in the experiment. We can use the sample space of elementary events to build up other sample spaces by collecting the elementary events with like characteristics.

You have probably played games, such as Monopoly and Parcheesi, in which your moves were determined by the throw of a die or a pair of dice. A sample space for the experiment of tossing a single die would be the set $\{1, 2, 3, 4, 5, 6\}$. It is rather more difficult to generate a sample space for the tossing of a pair of dice. To help us keep things straight, let us suppose that one die is red and the other one green. An outcome of 4 on the red die and 3 on the green die would be different from an outcome of 3 on the red die and 4 on the green die. If we agreed to write the outcome on the red die first and the outcome on the green die second, we could show this difference by setting down the pairs, (4,3) and (3,4).

TABLE 1

At the right is a sample space for the tossing of a pair of dice. It is assumed that one of the dice is red and the other green; also that the outcome of the red die is given first in each of the pairs shown in the table. Note that in the text the two outcomes, separated by a comma, are given in parentheses. Thus we would refer to (2,5), (4,3), (6,1) and so on.

| | | OUTCOME ON GREEN DIE | | | | | |
|--------------------|---|----------------------|--------|--------|--------|--------|--------|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| OUTCOME ON RED DIE | 1 | 1 1 | 1 2 | 1 3 | 1 4 | 1 5 | 1 6 |
| | 2 | 2 1 | 2 2 | 2 3 | 2 4 | 2 5 | 2 6 |
| | 3 | 3 1 | 3 2 | 3 3 | 3 4 | 3 5 | 3 6 |
| | 4 | 4 1 | 4 2 | 4 3 | 4 4 | 4 5 | 4 6 |
| | 5 | 5 1 | 5 2 | 5 3 | 5 4 | 5 5 | 5 6 |
| | 6 | 6 1 | 6 2 | 6 3 | 6 4 | 6 5 | 6 6 |

Table 1 shows the sample space for the experiment of tossing a pair of dice. Suppose that you tossed the dice and that the red die came up 4 (that is, came to rest with the number 4 uppermost) and the green die came up 3. Moving horizontally from 4, at the extreme left of the table, until we came to the column headed by 3, we would find the pair (4,3). We shall have occasion to refer to this sample space several times in the course of this article.

See if you can list sample spaces for the following experiments:

1. Toss a coin and a die. [Hint: two points in this sample space could be (H,3) and (T,2).]

2. Toss a nickel, dime, and quarter. If the nickel comes up heads, the dime tails, and the quarter heads, we can indicate this by (H,T,H). You should find 8 points for this sample space.

PROBABILITY OF AN EVENT

An *event* is by definition any subset of a sample space. In other words, if each member of set A is also a member of set B,

we say that A is a subset of B. If set A = {1, 2, 3} and set B = {1, 2, 3, 4, 5}, then set A is a subset of set B. Thus an event is a set of individual outcomes of an experiment. An elementary event, as we have indicated, is a single individual outcome. In order to assign probabilities to an event—that is, to a set of individual outcomes—we must first be able to assign probabilities to individual outcomes of the experiment.

If we consider, for example, the experiment of drawing a single card from a deck of cards, there are 52 elementary events or individual outcomes possible. If we make a random draw, each card has exactly the same chance of being drawn. Thus to each of the 52 possible outcomes we would assign the same probability— $1/52$. Note that the sum of the probabilities assigned to the elementary events is 1. The probability of an event that is certain is also 1.

The same basic principle may be used to answer the question "What are your chances of winning a sweepstakes contest or the door prize at a party?" In each case, if the total number of tickets is n and you hold only one ticket, then you would have

only one chance in n of winning. Thus the probability of your winning would be $1/n$.

In the experiment of drawing a card from a deck, we may be concerned only with whether the card is an ace. Then the event with which we are concerned consists of four elementary events—the ace of clubs, the ace of diamonds, the ace of hearts, and the ace of spades. We still would have to draw at random from the entire deck of cards, numbering 52. Since 4 out of the 52 cards belong to the event just described (4 aces), we would say that the probability that the outcome is an ace is $4/52$. We could use the symbols $P(\text{Ace}) = 4/52$ to represent this statement.

Suppose that in the above experiment we were concerned only with the color of the card chosen. Do you see why $P(\text{Red}) = 26/52$? We can use the probabilities assigned to elementary events to assign probabilities to the points in other sample spaces. For example, if we were using the sample space Clubs, Diamonds, Hearts, Spades, we could assign the probability of $1/4$ to each point in the sample space since $P(\text{Clubs}) = 13/52 = 1/4$.

We can summarize the above discussion by setting down the following rule:

If an experiment can result in n different but equally likely outcomes and if m of these outcomes correspond to event X , then the probability of the event is $P(X) = m/n$. (Rule 1)

Applying this rule, let us find the probabilities of some events in the experiment of tossing a pair of dice. Table 1 lists 36 possible outcomes for this experiment. Thus we could assign a probability of $1/36$ to each of the elementary events. Now consider the event “The sum of the numbers shown is 7.” If we let r stand for the number on the red die and g for the number on the green die, we can represent the event as $r + g = 7$. How many of the elementary events correspond to this amount? In other words, how many of the pairs in the table add up to 7? Consult the table to find the answer. You will see that $P(r + g = 7) = 6/36$. If we considered the event “The same

number appears on both dice,” we could call the event “ $r = g$ ” or, as is common in many games played with dice, “Double.” There are six such pairs—(1,1), (2,2), (3,3), (4,4), (5,5) and (6,6). Thus $P(r = g) = P(\text{Double}) = 6/36$. What is $P(r + g > 7)$? The symbol $>$ stands for “is greater than.” Are you able to find 15 points in the sample space that corresponds to this event? You should be able to, for $P(r + g > 7) = 15/36$.

These examples should suggest a way to answer the question “How likely are you to land on ‘Boardwalk’ in your next turn in a game of Monopoly?” Suppose, for example, that you are located on “Pennsylvania Avenue,” which is 5 spaces from “Boardwalk.” To determine the probability of your landing on “Boardwalk,” carry out the experiment of tossing two dice. You would land on “Boardwalk” if the numbers shown on the dice totaled 5. What is $P(r + g = 5)$?

Try the following problems:

1. If a box contains 4 red marbles and 5 white marbles, what is the probability of drawing a red marble on the first try?
2. What is the probability that you will draw a face card (Jack, Queen, King) from a deck of cards?
3. What is the probability that you will get a 5 when you toss a single die?
4. What is the probability that you will not get a 5 when you toss a single die?

COMPLEMENTARY EVENTS

There is a relationship which frequently simplifies the computing of a probability. Consider the example pictured in Figure 2. There are 10 points in the sample space S . The event A contains 4 of these points and the remaining 6 points are not in A . The set of points which are not in a given set A is called the complement of A and is usually indicated by the symbol \bar{A} . If the elementary events in the sample space S are equally likely, then $P(A) = 4/10$ and $P(\bar{A}) = 6/10$.

In a sample space of n equally likely events, if an event A occurs in m of the outcomes, then the event \bar{A} will occur in $n - m$ outcomes. Thus if $P(A) = m/n$, then $P(\bar{A}) = (n - m)/n$. Now $(n - m)/n = 1 - m/n$. We obtain this result by dividing both the nu-

merator and the denominator by n , as follows:

$$\frac{\frac{n-m}{n}}{\frac{n}{n}} = \frac{1 - \frac{m}{n}}{1} = 1 - \frac{m}{n}$$

We can now give the following general rule:

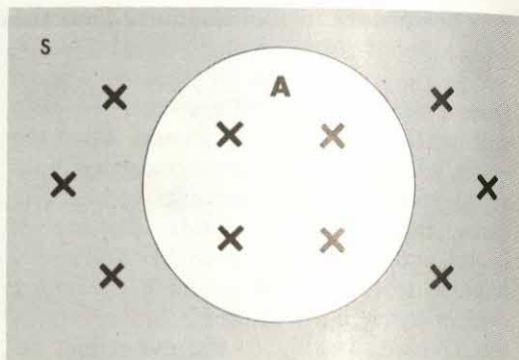
$$P(\bar{A}) = 1 - P(A) \quad (\text{Rule 2})$$

Sometimes it is easier to compute the probability of one of two complementary events than it is to compute the probability of the other. In such cases, we compute the easier probability. We then use the relationship indicated in Rule 2 to derive the probability of the complementary event. For example, in some games played with dice, there is either a premium or a penalty associated with throwing doubles. We could compute the probability of not throwing a double by counting the sample points in Table 1 which are not doubles. (There are 30 of them.) Or we could compute the probability of throwing a double ($R = G$) and use the relationship in Rule 2 to derive the probability of not throwing a double ($R \neq G$). The symbol \neq stands for "is not equal to."

PROBABILITY OF "A OR B"

A situation that often arises is the finding of the probability of an event that might be expressed as "either event A or event B." By the event A or B we mean the set of outcomes which correspond either to event A or event B or possibly both A and B.

Consider the example shown in Figure 3. There are 10 points in the sample space S. In this sample space, $P(A) = \frac{3}{10}$ and $P(B) = \frac{2}{10}$. Since the event A or B contains the 3 elements of A and the 2 elements of B, then $P(A \text{ or } B) = P(A) + P(B) = \frac{5}{10}$. In this instance, there are no elements that are in both A and B. A and B here are mutually exclusive events, meaning that they cannot both occur in a single trial of an experiment. When events A and B are mutually exclusive, we can find the probability of A or B simply by adding the probability of A to the probability of B as follows:



2. Diagram illustrating complementary events.

If events A and B are mutually exclusive, then $P(A \text{ or } B) = P(A) + P(B)$ (Rule 3)

For example, suppose we were required to find the probability of throwing either a 7 or a double in a toss of two dice. Since 7 is an odd number, it is not possible for a single toss of two dice to produce both a 7 and a double. Thus the event $(r + g = 7)$ and the event $(r = g)$ are mutually exclusive. We have already seen that $P(r + g = 7) = \frac{6}{36}$ and that $P(r = g) = \frac{6}{36}$. Hence, applying rule 3, $P(r + g = 7 \text{ or } r = g) = \frac{6}{36} + \frac{6}{36} = \frac{12}{36}$. Can you verify this by counting the appropriate points in the sample space shown in Table 1?

When both events are not mutually exclusive—that is, when they can both occur in a single outcome—we cannot use the addition principle of Rule 3. Consider the situation presented in Figure 4. In this example, the sample space S consists of 10 equally likely elementary events; 5 outcomes correspond to event A and 4 outcomes to event B. Hence $P(A) = \frac{5}{10}$ and $P(B) = \frac{4}{10}$. If we were to apply Rule 3 here, we would have $P(A \text{ or } B) = P(A) + P(B) = \frac{5}{10} + \frac{4}{10} = \frac{9}{10}$. This answer would be incorrect. The reason is that events A and B are not mutually exclusive, since 2 outcomes belong to both A and B. To obtain a correct answer in this case for $P(A \text{ or } B)$, we need to apply the following rule:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \quad (\text{Rule 4})$$

Let us consider an example in which this rule could be used. Suppose we wanted to find the probability of drawing either a face card or a spade from a deck of cards. $P(\text{Spade}) = \frac{13}{52}$ and $P(\text{Face Card}) = \frac{12}{52}$. (Remember that each of the four suits has three face cards—Jack, Queen, and King.) There are three cards which are both face cards and spades. Therefore $P(\text{Spade and Face Card}) = \frac{3}{52}$. Applying Rule 4, we would conclude that $P(\text{Spade or Face Card}) = \frac{13}{52} + \frac{12}{52} - \frac{3}{52} = \frac{22}{52}$. Could you list the 22 cards that would belong to the event “Spade or Face Card?”

If we agree that the probability of an event which cannot occur is 0, then Rule 4 could replace Rule 3, involving mutually exclusive events. This is why. When two events A and B are mutually exclusive, then the event A and B cannot occur and we would say $P(A \text{ and } B) = 0$. Applying Rule 4, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = P(A) + P(B) - 0 = P(A) + P(B)$. This would be the equivalent of Rule 3: $P(A \text{ or } B) = P(A) + P(B)$.

Here are two rather simple problems for you to try:

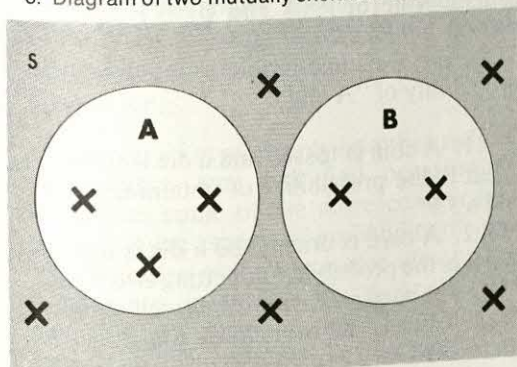
1. What is the probability that you could throw either a 7 or an 11 on a single throw of 2 dice?

2. What is the probability that you would throw either a double or a total of more than 9 on a single toss of 2 dice?

PROBABILITY OF “A AND B”

When two events, A and B, are independent, we can compute the probability of the event A and B by using the following

3. Diagram of two mutually exclusive events.



rule, involving the multiplication principle:

If events A and B are independent, then $P(A \text{ and } B) = P(A) \cdot P(B)$ (Rule 5)

Intuitively, we would expect two events to be independent when they have nothing to do with each other. Although this is usually a reliable rule of thumb, it must be used with care. For example, when two events are mutually exclusive, our first thought might be that they have nothing to do with each other. But when one of two mutually exclusive events occurs, the other cannot possibly occur. Thus the occurrence of one of these events certainly affects the probability that the other also occurs. Mutually exclusive events, therefore, are never independent of one another. For two events to be independent, the occurrence of one must not affect the probability that the other occurs at the same time.

Here is a problem involving two independent events, in which Rule 5 could be applied. Suppose that we toss two dice. What is the probability that the number 1 would come up on the red die and that the number 3 would come up on the green die? In this case, the outcome on the green die has nothing to do with the outcome on the red die. We therefore have two independent events. The possibility of any one of the numbers 1, 2, 3, 4, 5 and 6 coming up on the toss of a single die is $\frac{1}{6}$. The probability of the outcome (1,3) on a toss of two dice would be calculated as follows, in accordance with Rule 5: $P(1 \text{ and } 3) = P(1) \cdot P(3) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$.

By using the multiplication principle, we can provide an answer of sorts to the question “If a baseball player averages 3 hits in 10 times at bat, what are his chances of getting 4 straight hits?” If we assume that the times at bat are independent trials, we could compute the possibility of 4 straight hits in this way: $P(\text{HHHH}) = P(H) \cdot P(H) \cdot P(H) \cdot P(H) = \frac{3}{10} \cdot \frac{3}{10} \cdot \frac{3}{10} \cdot \frac{3}{10} = \frac{81}{10,000}$. We would expect the batter to get 4 hits in a row about 81 times in every 10,000 sequences of 4 times at bat. This probability would hold if we were right in assuming that the 4 times at bat would be

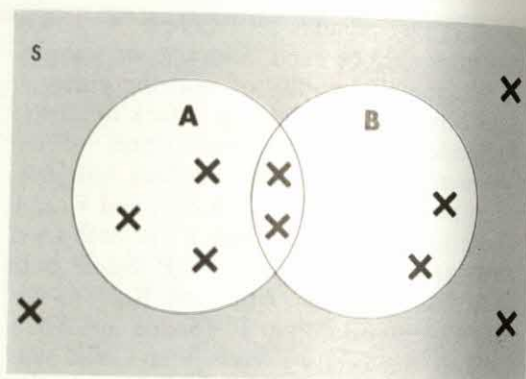
independent trials. The assumption might not be a sound one in this case. For one thing, the batter would probably hit better against certain pitchers than against others. Also, getting a hit or not getting a hit might have an effect on his chances the next time he came to bat.

Consider the event "Double and Even" in the experiment of tossing two dice. By "Double," of course, we mean that the same number would come up for both dice; by "Even," that the sum of the two numbers would be an even number. In this case, $P(\text{Double}) = \frac{6}{36}$ and $P(\text{Even}) = \frac{18}{36}$. If we applied the multiplication principle for independent events, as stated in Rule 5, we would have $P(\text{Double and Even}) = \frac{6}{36} \cdot \frac{18}{36} = \frac{108}{1,296} = \frac{3}{36}$. But if we examine Table 1, we will find 6 points that correspond to the event Double and Even. Therefore the true value of $P(\text{Double and Even})$ is $\frac{6}{36}$ and not $\frac{3}{36}$, as our calculation had seemed to indicate. The reason is that in this case the two events are not independent. If we throw a double, we are certain to throw an even number. This means that if we have thrown a double, the probability that we have also thrown an even number is 1. Again, if we know that we have thrown an even number, the probability that we have also thrown a double is $\frac{1}{3}$. This is determined by counting the number of ways an even number can occur (18 in all) and the number of these occurrences which are doubles (6).

In order to calculate $P(A \text{ and } B)$ when the events are not independent, the multiplication principle for independent events (Rule 5) can be generalized as follows:

$$P(A \text{ and } B) = P(A) \cdot P(B/A) \quad (\text{Rule 6})$$

B/A would be read as "B given A." When we know that the event A occurs, we can compute $P(B/A)$ by thinking of the event A as a reduced sample space. To show what I mean, examine Figure 5. The sample space S consists of 10 equally likely events, of which 5 are in A, 3 are in B, and 2 are in both A and B. Given that A occurs, then any outcome obtained is one of the 5 in A.



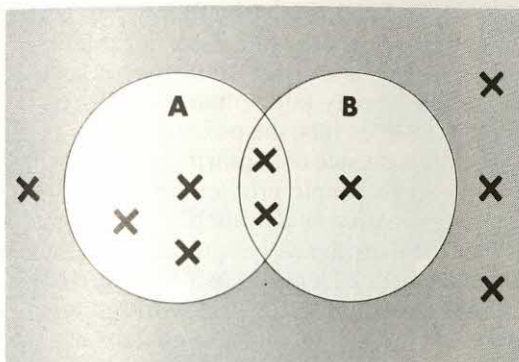
4. Diagram of events not mutually exclusive.

If we think of these 5 points as a new sample space, then 2 of the equally likely outcomes belong to the event B. Thus $P(B/A) = \frac{2}{5}$. Applying Rule 6, we can compute $P(A \text{ and } B) = P(A) \cdot P(B/A) = \frac{5}{10} \cdot \frac{2}{5} = \frac{10}{50} = \frac{2}{10}$. This, as you can see, agrees with what we would obtain by a direct count of the points that belong to both A and B in Figure 5.

If events A and B are independent, then the probability that B occurs will not be affected by the fact that A occurs. Thus for independent events, $P(B/A) = P(B)$. Suppose we were to select one card from a deck of cards. What is the probability that the card selected will be both a red card (R) and a face card (F)? Since half the cards are red, we know that $P(R) = \frac{26}{52}$. Of the 26 red cards, 6 are face cards. Hence $P(F/R) = \frac{6}{26}$. Since $P(F) = \frac{12}{52} = \frac{6}{26}$ and $P(F/R)$ is also $\frac{6}{26}$, the events F and R are independent and $P(R \text{ and } F) = P(R) \cdot P(F)$. Thus the multiplication rule for independent events (Rule 5) is a special case of the more general Rule 6: $P(A \text{ and } B) = P(A) \cdot P(B/A)$.

Here are some problems involving the probability of "A and B."

1. A coin is tossed and a die is thrown. What is the probability of obtaining heads and a 3?
2. A card is drawn and a die is tossed. What is the probability of getting two 6's?
3. A card is drawn from a well-shuffled deck. What is the probability that it is neither a face card nor a red card?



5. Diagram illustrating reduced sample space.

DEGREE OF CONFIDENCE

In this brief introduction to probability, we have touched on only a few of the basic principles involved. The scope of probability theory is much broader than our discussion would suggest. For example, we simplified matters a good deal by dealing only with situations in which the individual outcomes of an experiment were equally likely. In many situations, this is not the case. Calculations then become more complex and new dimensions are added to the study of probability.

In our simplified presentation, we merely pointed out that predictions based on probability are uncertain. The scientist accepts this, but he wants to know how uncertain. He recognizes that in trying to predict the outcome of an experiment subject to the laws of chance, he may often be wrong. He needs to know how often and how wrong. He wants to establish the degree of confidence he can place in his predictions. Probability theory provides the basis for establishing this particular degree of confidence.

MANY USES

If you go at all deeply into the study of mathematics, you will find occasion to work with some of the more complicated and intriguing aspects of probability theory. You will also obtain a clearer idea of the many ways in which this theory serves man. Let me point out here a few of its uses. It enables scientists to fix a limit with-

in which the deviations from a given physical law must fall if these deviations are not to count against the law. It has been used to calculate the positions and velocities of electrons orbiting around the nuclei of atoms. The fluctuations in density of a given volume of gas have been analyzed by applying probability theory. It has played an important part in genetics; among other things, it has made it possible to calculate the percentage of individuals with like and unlike traits in successive generations. Manufacturers use probability theory to predict the quality of items coming off mass-production lines. Insurance experts make extensive use of the theory. It enables them, for example, to calculate life expectancies so that they may set appropriate life insurance rates. Finally, because of probability theory, electronic computers can be programmed to predict the outcome of elections on the basis of comparatively few returns and with what is generally a surprising degree of accuracy. The comparatively few failures of computerized election prediction cannot obscure the success attained by this technique.

Blaise Pascal, who developed the theory of probability.

Bettmann Archive



GAME THEORY

by Joseph G. Cowley

Five criminals huddle in a building, plotting a crime. Outside waits a lone policeman, determined to capture the leader. The criminals don't know he is out there. Nonetheless, they plan to leave randomly, one at a time, to avoid drawing attention to themselves. All the policeman knows is that the leader is the tallest of the criminals. He cannot capture more than one of the men, for he would be overpowered. He has no way of getting help. As the men come out, which one should the policeman arrest? The first? The second? The last?

If he made the arrest on a random basis, the policeman would only have a 20 per cent chance of getting his man. Game theo-

rists say he can do better by following a certain strategy, or plan of action: let the first two men go, and arrest the next man to come out who is taller than the first two men. If he does this, the policeman stands a 40 per cent chance of capturing the leader.

This is a simple problem in game theory, a fascinating new branch of mathematics that deals with risk in conflict situations. It gets its name from the fact that so many conflict situations in the real world of business, finance, the military, and so on are similar in basic structure to the parlor games we know so well: bridge, poker, checkers, chess, even ticktacktoe. Both in games and in everyday events, people compete with one another. There are rules by which the game must be played; there are outcomes, or payoffs—such as win, lose, and draw—that result from the opponents' different moves, or strategies; and there is generally some information available to one or more of the players involved.

In game theory it is assumed that the

Two boys playing chess, a zero-sum game involving a pair of players following set rules, with a clear objective—simply to win.

Mimi Forsyth, Monkmeier





Jan Braunle



Mickey Palmer, DPI

The complicated game of football, involving two teams pitted against each other to win a game, with a higher numerical score. There are rules; but all kinds of strategies and tactics can be worked out. Left: the coach explains a play to members of the team. Right: a play in action.

objective of each player is to maximize his gains or minimize his losses. It is also assumed that he is faced with a rational opponent whose objective is to do the same. Of course, the cop-and-robbers situation is really a one-person game. The robbers don't know they are competing with the policeman. If they did, they would try a strategy of their own. They might assume that the policeman is a game theorist and therefore send their leader out first. On the other hand, the policeman might realize that they would think this, and therefore arrest the first man to come out. In short, a game can be complicated.

Aside from one-person games, which are apt to have little value and to be easy to solve, perhaps the simplest games are games of two persons (individuals or teams) where the losses of one are the gains of the other. This kind of game is called a two-person zero-sum game. Zero-sum means that what one side gains, the other side loses—and vice versa. The total value of the game doesn't increase or decrease during the course of playing the game.

It is easy to imagine games where this

isn't the case. Two advertisers competing for a market, for example, might not only increase or decrease their shares of the market through their advertising, but might actually increase the total market. To best illustrate game theory, we will stick to the two-person zero-sum game.

CASE OF THE COMPLICATED COLLECTION

Most games are finite. That is, there are only a certain number of strategies, or

FIGURE 1.

Game Matrix for You and Oscar

| | | Oscar's strategies | | |
|-----------------|---|--------------------|----|-------------------|
| | | A | B | |
| Your strategies | A | 30 | 60 | (30) Your maximin |
| | B | 40 | 20 | |
| Oscar's minimax | | (40) | 60 | |

alternatives, that each player can follow and the game is “solved” in a certain number of plays, or moves. Some games are also games of perfect information: we know, or can see, each move our opponent makes. Checkers, chess, and ticktacktoe are examples of finite games of perfect information. Chess tends to be interesting because so many different moves are possible. A game like ticktacktoe tends to be less interesting because there are so few moves and it can be solved so easily. In fact, in ticktacktoe the first player to go can assure himself of at least a draw if he makes his mark in one of the corner squares. The second player can guarantee a draw if he makes his mark in the center square—assuming, of course, that each subsequent move is made to block his opponent.

To illustrate game theory, let's consider a finite, two person zero-sum game of imperfect information. Let's assume that Abe owes you \$60, and Bill owes you \$40. To collect the money you must be at either Abe's or Bill's place of employment on payday at 5 P.M. The situation is further complicated by the fact that these two men owe the same amounts of money to a fellow we will call your opponent: Oscar.

Whoever is at Abe's or Bill's place of employment at 5 P.M. gets the money. If

both of you show up at one place, you split the money. But if neither of you shows up, Oscar gets the money because he is a special friend of both Abe and Bill. Both you and Oscar have this information. Thus you must determine where you should go to maximize your gain: to Abe's or Bill's?

This is a problem that can be solved by using game theory. It is a conflict situation: you want to get as much money as you can while Oscar, who could get all the money if you didn't go to either place, wants to minimize his losses. It is a zero-sum game because the total amount to be paid out (\$100) stays the same whatever strategies are used.

CHOICE OF STRATEGIES

The game you and Oscar play is called a 2×2 game: there are 2 strategies available to you and 2 available to Oscar. Each of you can go to see either Abe (strategy A) or Bill (strategy B). Some two-person games have many more strategies available to the players. For example, in a 2×3 game, your opponent has 3 strategies to choose from while you have only 2. In a 4×3 game, you have 4 strategies to your opponent's 3. In a $3 \times m$ game, m means that your opponent has many strategies from which to choose.

Paul Conklin, Monkmeyer



A shopper studying brands of products in a grocery. Manufacturers and advertisers apply game theory to capture a share of the food market.

Large games can usually be reduced to more-manageable proportions because some of the strategies will not make sense in terms of the game. For example, if all the payoffs for a particular strategy are zero, there is no reason to choose this strategy. It may as well be eliminated from consideration.

We can best illustrate a game by a *payoff matrix*. This indicates what happens when the players select their different strategies. Figure 1 shows a payoff matrix for the game between you and Oscar. This matrix shows your payoffs when you and Oscar select either of your different strategies.

It is conventional to have a matrix show payoffs for the player whose strategies are listed on the left side of the matrix. This is the player who wants to maximize his gains. The payoffs shown are gains, in a zero-sum game, for the player whose strategies are listed at the left. The payoffs are losses for the person who is named at the top of the matrix.

The matrix shows that if both you and Oscar choose strategy A (going to see Abe), your payoff will be \$30. If you choose A and Oscar chooses B (going to see Bill), your payoff will be \$60. If you choose B and Oscar chooses A, your payoff will be \$40. If you both decide to choose strategy B, your payoff will be \$20.

Some additional figures are shown outside the squares in Figure 1. These help us "solve" the problem following game-theory procedure. The particular theory we will follow was first proposed by the mathematician John von Neumann in 1929 and extensively developed in a book he wrote with Oskar Morgenstern, published in 1944, titled *Theory of Games and Economic Behavior*.

The figures outside the matrix are called *rim figures*. The column at the right gives the row minima: the minimum amount in the row opposite which it appears. The row at the bottom of the matrix gives the column maxima: the maximum amount in the column under which it appears. Game theory says that you should choose the strategy that provides the maxi-

mum of your minimum gains. This is called your *maximin*. Your opponent, if he is rational, should choose the strategy that provides the minimum of his maximum losses. This is called his *minimax*. In Figure 1, your maximin and Oscar's minimax are circled.

The purpose of game theory is to eliminate chance or probability from your problem. By choosing strategy A you guarantee yourself a gain of at least \$30. If your opponent is stupid enough to choose strategy B, you'll make \$60. Oscar, on the hand, will choose strategy A because it guarantees him a maximum loss of \$40—if you try to be wily and choose strategy B. But Oscar knows you won't try strategy B, since you would make only \$20 if he, too, chooses B. Either one of you could, of course, try to outsmart the other. But then you would be gambling, and this is what game theory avoids. It is, instead, a mathematical technique for guaranteeing a minimum gain (or maximum loss) from a conflict situation.

MIXED STRATEGIES AND SADDLE POINTS

We assumed that your game with Oscar would be played only once. That is, each player was to make only one move. But most games, in both real life and among parlor games, involve a number of moves. Matching pennies is a good example. In this game, if you choose a "pure" strategy, such as playing heads all the time, you could wind up losing your shirt (or at least all your pennies). Frequently, such a situa-

FIGURE 2.
Game Matrix with Saddle Point

| | | Oscar's strategies | | |
|-----------------|---|--------------------|-----|------|
| | | A | B | |
| Your strategies | A | 50 | 100 | (50) |
| | B | 40 | 20 | 20 |
| | | (50) | 100 | |

tion calls for a “mixed” strategy, like playing heads and tails each 50 per cent of the time on a random basis. Tossing the coin before each play would achieve this aim. In fact, even those who know nothing about game theory usually play the game this way.

Now let’s look at your game with Oscar again and assume that it is to be played many times. Does a pure strategy then make sense? Not quite. We’ve determined that Oscar will choose strategy A because, if he doesn’t, you might get \$60 instead of \$30. But let’s fool him. Let’s choose strategy B, getting \$40 instead of \$30. How long will Oscar allow this? If you choose B, then he will also choose B, thus reducing your winnings to \$20. You can switch back to strategy A and gain \$60. But he’ll also switch back to strategy A. Before long, both of you are employing mixed strategies, trying to best each other. Now the question is: what kind of mixed strategy should you employ?

Before we answer this question, let’s consider whether or not a mixed strategy should be used in a particular game. Game theory says that a mixed strategy should not be used if the game has a *saddle point*. A game is said to have a saddle point if the maximin equals the minimax; that is, if the maximum of your minimum gains equals the minimum of your opponent’s maximum losses. The game with Oscar does not have a saddle point, as you can see by comparing the two circled figures in Figure 1. Therefore, the game calls for a mixed strategy. If the two circled figures were equal, then the game would have a saddle point, and each player should choose the pure strategy dictated by his maximin or minimax. Let us illustrate this.

In Figure 2 we have changed the pay-offs (or payouts, as they are sometimes called) to create a saddle point for the game. Your maximin equals Oscar’s minimax. Under these conditions it doesn’t make sense for you to choose strategy B. You would only lose. Oscar, who assumes that you are rational, cannot do other than choose strategy A, too. If he chooses strategy B, he stands to lose \$100 instead of

FIGURE 3.

Game Matrix with Oddments

| | | Oscar's strategies | | |
|-----------------|---|--------------------|----|----|
| | | A | B | |
| Your strategies | A | 30 | 60 | 30 |
| | B | 40 | 20 | 20 |
| | | 10 | 40 | |

\$50. So both of you will continue to choose strategy A. Such a game is said to be strictly determined.

Now let’s return to the original game and see what mixed strategy each player should employ to maximize his gains (assuming that the game is played many times).

VALUE OF THE GAME

Every game has a certain value. The value of the strictly determined game in Figure 2 is \$50. The value of the original game (in Figure 1), played on a one-time basis, is \$30. (Both these values are, of course, yours, and we will continue to speak of the value of the game in your terms.)

If the original game is to be played on a continuing basis, both you and Oscar will begin mixing your strategies, each trying to get the better of the other. What the two of you are really fighting over is the \$10 spread between your guaranteed minimum gain of \$30 and his guaranteed maximum loss of \$40—if each of you plays a pure strategy. By mixing strategies each of you hopes to get as much of this \$10 as possible. This suggests that the value of the game lies somewhere between \$30 and \$40. How can Oscar assure himself that you get no more than that amount? The answer lies in mixing strategies in a prescribed way.

After determining that the game does not have a saddle point, erase the rim figures. Determine the absolute differences between the payoffs in each row and column and write these figures in the margins. These figures are called *oddments*, or part of the odds in using each of the strategies.

FIGURE 4.

Game Matrix for Prisoners' Dilemma

| | | Prisoner B | |
|------------|----|------------|------|
| | | C | NC |
| Prisoner A | C | 5,5 | 0,20 |
| | NC | 20,0 | 1,1 |

We've done this for you in Figure 3. You can see that the differences in the row payoffs are \$30 and \$20, while the differences in the column payoffs are \$10 and \$40. These figures are the ratios with which the different strategies should be played (3 to 2 for you, and 4 to 1 for Oscar). Each figure, however, applies to the opposite strategy. In other words, your mixed strategy should be to play strategy B 3 times and strategy A twice out of every 5 plays; Oscar's mixed strategy should be to play strategy A 4 times and strategy B once in the same number of plays.

Even though your opponent can easily figure out your overall strategy, it is best to keep him guessing about each individual move. Otherwise he might outfox you. You should choose each move on a random basis. For example, you might throw 3 red cards, representing strategy B, into a hat together with 2 black cards representing strategy A. Before each move mix the cards and withdraw one. Let this card determine the strategy you choose. Throw the card back into the hat and repeat the process for the next move. In the long run you will play your strategies in a ratio of 3 to 2, and the probability is high that you will achieve the value of the game. The value of the game is easily determined. For each row, multiply the oddment times each payoff in the other row. Add the products together and divide by the sum of the oddments. Average your answers for the rows. This gives the value of the game if you play your prescribed mixed strategy. Do the same for the columns. This gives the value of the game if your opponent plays his prescribed mixed strategy. These two values should be the

same. Over the long haul, the value of the game will be achieved if both of you play your prescribed mixed strategies. If one player deviates from his prescribed strategy, chances are he will not achieve the value of the game.

The arithmetic involved in determining the value of the game is illustrated in these equations:

$$\frac{20 \times 30 + 20 \times 60}{20 + 30} = 36$$

$$\frac{30 \times 40 + 30 \times 20}{20 + 30} = 36$$

$$\text{Value of the game} = \frac{36 + 36}{2} = 36$$

$$\frac{40 \times 30 + 40 \times 40}{40 + 10} = 56$$

$$\frac{10 \times 60 + 10 \times 20}{40 + 10} = 16$$

$$\text{Value of the game} = \frac{56 + 16}{2} = 36$$

The value of the game, if both players play rationally (i.e., use their prescribed mixed strategies), is \$36.

NON-ZERO-SUM AND N-PERSON GAMES

When we move away from the two-person, zero-sum game, problems become a bit more complicated. For one thing, psychology, negotiation, and communication may become factors in the problems.

A non-zero-sum game, you will recall, is one in which the losses of one player are not necessarily the gains of the other. That is, the total value of the game does not necessarily remain the same throughout the play. An example of a two-person, non-zero-sum game is "the prisoners' dilemma." lemma."

Imagine that two criminals suspected of a bank robbery are arrested and placed in separate jail cells so they cannot communicate. If one confesses and turns state's evidence, he will go free. The other will be sentenced to 20 years in prison. If both criminals confess and throw themselves on the mercy of the court, they will each receive a 5-year sentence. If neither con-

feesses, they will each get 1 year for carrying concealed weapons. If you were one of the prisoners, what would you do?

This conflict situation is pictured in Figure 4. The strategies are confession (C) and no confession (NC). Because this is a non-zero-sum game, the payoffs for both players are listed in each square, with player A's payoff to the left of the comma and player B's payoff to the right. Study only your own payoff. As a good game theorist, you should choose the strategy that provides your minimax, the minimum of your maximum losses. Thus you should confess. The worst that could happen would be a 5-year sentence.

Your opponent, prisoner B, should choose the same strategy. But this may be oversimplifying the game. Suppose you think your opponent is someone who would never confess. What would you do then? It depends on what type of person you are.

Suppose you two were cellmates and could discuss the situation. Would you both agree not to confess? That would make sense. But what about the danger of a double-cross?

As you can see, for many (if not most) games, mathematics doesn't have all the answers.

Games involving more than two people (n-person games) can be even more complex. Imagine a situation involving

three people: A, B, and C. If A cooperates with B they will split \$6. If A cooperates with C they will split \$8. If B cooperates with C they will split \$10. This is illustrated in the triangle in Figure 5. Who should cooperate with whom to divide the spoils?

"Look," A says to B, "cooperate with me and I'll give you \$4 and only keep \$2."

"Don't be silly," says B to A. "I can probably get \$5 cooperating with C."

"Nothing doing," says C. "I think I can make a better deal with A. How about it, A? Will you take \$2 and leave me \$6?"

"Okay," says A.

B panics. "Look, A," he says, "I'm willing to split the \$6 fifty-fifty."

"Don't lose your head," says C to B. "I'll give you \$4."

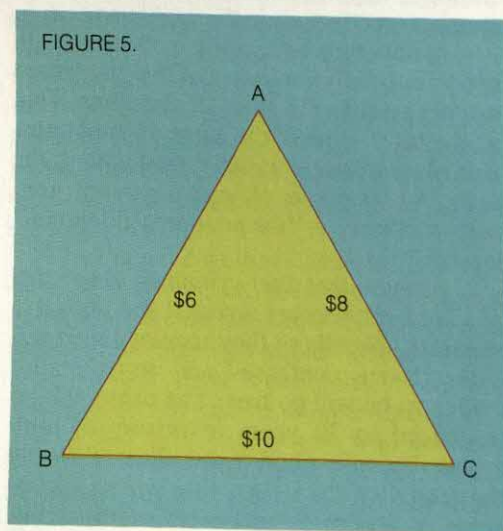
"But that's what I offered you," says A.

Round and round it goes. The division of the money will depend on the negotiating talents of those involved. All game theory can tell us is that, by cooperating with one of his opponents, the most A can hope to receive is \$2; the most B can hope to get is \$4; and the most C can hope to receive is \$6.

But let's change the rules. Suppose A, B, and C could gain a total of \$18 if all three of them cooperated. How should the money be split? The resolution above suggests that A should receive $\frac{1}{6}$, or \$3; B should receive $\frac{1}{3}$, or \$6; and C should receive $\frac{1}{2}$, or \$9.

As you can see, these simple games give only a hint of the complexity of larger, non-zero-sum games. Game theory has had its greatest success to date with only the most elementary of games, or conflict situations. Undoubtedly it will come to have increasing importance in the future, for so much of life has to do with conflict and competition. As John D. Williams has written in his book *The Compleat Strategist*: "The concept of a strategy, the distinctions among players, the role of chance events, the notion of matrix representations of the payoffs, the concepts of pure and mixed strategies, and so on give valuable orientation to persons who must think about complicated conflict situations."

FIGURE 5.



CALCULUS

by Murray Spiegel

One of the greatest contributions to modern mathematics, science, and engineering was the invention of *calculus*, or, as it is sometimes called, *the calculus*, near the end of the 17th century. It is safe to say that without this fundamental branch of mathematics many technological accomplishments, such as the landing of men on the moon, would have been very difficult, or impossible, to achieve.

The word "calculus" comes from the Latin word for pebble. This name probably originated because pebbles were used thousands of years ago for counting and doing problems in arithmetic. Similar words that we often use are *calculate*, *calculation*, and *calculator*.

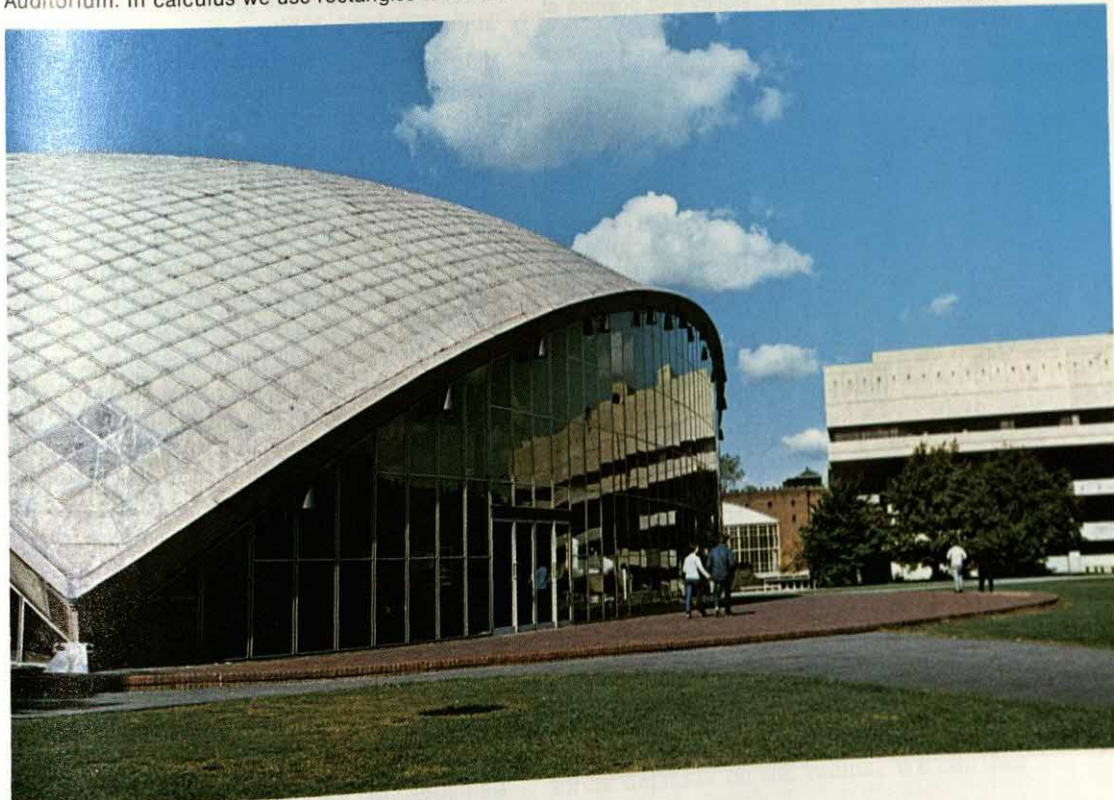
Two people who lived during the 17th century are credited with the invention of the calculus: Sir Isaac Newton of England, and Baron Gottfried Wilhelm von Leibniz of Germany. The basic ideas of calculus were developed independently by them within a few years of each other.

Newton, who was one of the greatest physicists of all time, applied the calculus to his theories of motion and gravitation. These theories, often referred to as *Newton's laws*, enabled him to describe mathematically the motion of all objects in the universe from the tossing of a ball into the air to the revolution of the earth, and the other planets of the solar system, around the sun.

Before Newton and Leibniz, the mathematics used for solving problems was the kind commonly taught in modern secondary schools. This involved subjects such as arithmetic, algebra, geometry, and trigo-

Nearly rectangular windows fill out the area under the roof of M.I.T.'s Kresge Auditorium. In calculus we use rectangles to find the area under a curve.

Calvin Campbell



nometry. The basic principles of these subjects were known at least 1,500 years before Newton and Leibniz. Although the mathematical principles studied in these subjects were useful in solving certain kinds of problems, they were not at all suited to solving problems dealing with *changing, or varying, quantities*. It was for the purpose of working with changing, or varying, quantities in everyday life that the calculus was invented. We can therefore say that calculus is the *mathematics of change*.

CHANGING, OR VARYING, QUANTITIES

If a man travels in an automobile at a velocity of 50 kilometers per hour, we know that in 2 hours he will travel a distance of 100 kilometers. In practice, of course, a driver rarely travels at the same velocity of 50 kilometers per hour for 2

Scientists use calculus to calculate the "escape velocity" and the rocket's velocity and position at any given time.

NASA



hours. The driver will sometimes stop, sometimes travel at a velocity of 80 kilometers per hour and sometimes at 40 kilometers per hour.

The velocity of the automobile is usually a changing, or varying, quantity. When the velocity is increasing, as when the auto goes from 30 to 40 kilometers per hour, we often say that the automobile is being *accelerated* or that it is undergoing *acceleration*. Similarly when the velocity is decreasing, as when the auto goes from 40 kilometers per hour to 20 kilometers per hour, we often say that the automobile is being *decelerated*, or undergoing *deceleration*. If the automobile maintains the same velocity, then we say that it is traveling at a *constant, or uniform, velocity*.

Another example of a changing quantity involves a ball that has been dropped or thrown. Suppose we drop a ball from a building. At the instant we drop it, the velocity is zero. Gradually the velocity increases; that is, the ball *accelerates*. Finally when the ball hits the ground, it is traveling at its greatest velocity.

Similarly if we throw a ball up into the air, it at first travels fast—that is, with a large velocity. Gradually it slows down, or *decelerates*, until its velocity is zero and it stops for an instant. At this point the ball has reached its maximum height and starts to come down. As it comes down, its velocity increases until the ball hits the ground. The change in the velocity of the ball—or any object that is thrown—is due to the attraction of the earth: the force of gravity, or the gravitational force.

The same ideas apply to a rocket launched from the earth's surface. By using calculus, it is possible to find the velocity the rocket must have in order to escape the earth's gravity—that is, not return to the earth. This velocity, often called the *escape velocity*, turns out to be about 11 kilometers per second, or about 40,000 kilometers per hour. By means of calculus we can calculate the time it would take for the rocket to get to the moon, and how much fuel would be needed.

There are many other examples of changing, or varying, quantities. When a



Tony Duffy

This car's velocity is hardly ever constant. The car accelerates at the start of the race, slows down before each curve, and accelerates again on straightways.

raindrop or snowflake falls, its size gradually increases. The population of a country changes each year, or even each day. The cost of living changes. The amount of a radioactive substance, such as radium, changes.

WHAT IS A VARIABLE?

Any quantity that is changing, or varying, is often referred to in mathematics as a *variable*. In the examples of the automobile, ball, and rocket, the velocity is a variable. Also the distance of the ball or rocket from the earth's surface is a variable. Even the time, which is changing, is a variable. The population of a country or the cost of living is a variable. When we put air in our tires the air pressure changes and is a variable. The outdoor temperature and humidity are variables.

In mathematics we often represent variables by symbols, such as letters of the alphabet. Thus we can let v represent velocity, t represent time, p represent pressure in a tire, and so on. These letters stand for numbers. Thus when the velocity is 50

kilometers per hour we can say that $v = 50$. When it is 25 kilometers per hour we can say that $v = 25$, and so on.

In the case of time, we usually measure it from some specific instant, which is often called the *time origin* or *zero time*. For example, the instant at which we drop a ball from a building is zero time, or $t = 0$. After 3 seconds have elapsed, we say that $t = 3$, after 5 seconds, $t = 5$, and so on.

Very often, in practice, we find that one variable depends in some way on another variable. For example, the distance a rocket or ball travels depends on the *time* of travel. We call distance the *dependent variable*. Time is called the *independent variable*.

In general, time is considered as an independent variable, while any variable that depends on it is a dependent variable. Thus the cost of living, which depends on time, is a dependent variable. The outdoor temperature is also a dependent variable.

Other independent variables besides time can occur. For example, the area of a circle depends on the radius. We can call

the area A the dependent variable, and the radius r the independent variable.

WHAT IS A FUNCTION?

If one variable depends on another, mathematicians say that the first variable is a *function* of the second variable. The distance a rocket or ball travels is a function of the time of travel. The area of a circle is a function of the radius of the circle.

Suppose we designate by the letter x any independent variable such as time, radius of a circle, and so on. Suppose further that we designate by the letter y any dependent variable that depends on x , such as the distance traveled by a rocket, or the area of a circle. Then the statement that y *depends on* x or y is a *function of* x is often abbreviated by mathematicians as $y = f(x)$. This is read as y *equals* f of x . The letter f is used as the abbreviation for *function*.

In terms of this functional notation, we could abbreviate the statement that the cost of living C is a function of the time t by writing $C = f(t)$. Similarly we could express the fact that the area A of a circle is a function of the radius r by writing $A = f(r)$.

A function indicates that there is a relationship between the dependent and independent variables. Thus $C = f(t)$ indicates that there is a relationship between the cost of living C and the time t . Similarly $A = f(r)$

indicates that there is a relationship between the area A of a circle and its radius r . One of the important problems in mathematics and its applications to other fields, such as science, engineering, and economics, is to determine the nature of the relationships between variables.

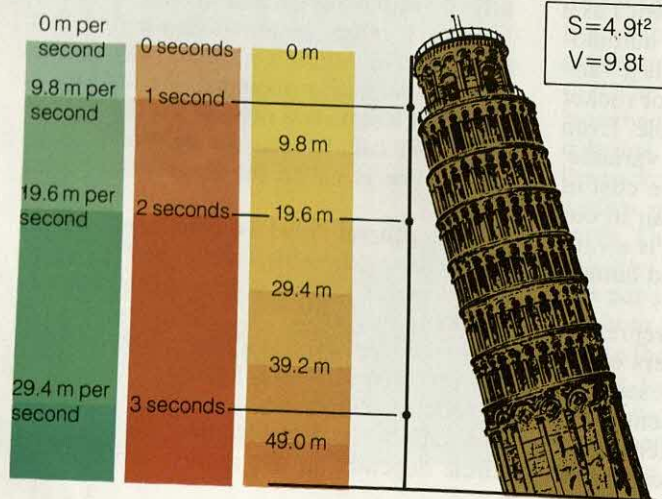
In some cases the relationship between variables is simple. For example, we know from geometry that the area A of a circle is given in terms of its radius r by means of the relationship $A = \pi r^2$ where π is a constant whose value is $3.14159 \dots$, or approximately $22/7$.

In other cases the relationship between variables is very difficult to obtain. For example, the relationship between the cost of living and time is not known, although we can have some idea about such a relationship based on past experience. However, even though we cannot find such a relationship, it does not mean that there is none.

RATES OF CHANGE

Suppose that at 10:00 A.M. an automobile driver is 30 kilometers from a certain town, and 2 hours later he is 160 kilometers away. The change in distance in 2 hours is then $160 - 30$, or 130 kilometers. On dividing this change in distance by the change in time—that is 130 kilometers divided by 2

Velocity. Time. Distance.



Suppose a ball is dropped from the top of the Leaning Tower of Pisa. The diagram at left shows the velocity of the ball and the distance it has fallen at given times. Using calculus one can determine the velocity and distance traveled at any given time.

hours—we obtain 65 kilometers per hour. This is called the driver's *average velocity*. It should be noted that, during the 2 hours, the driver may have been traveling at 80 kilometers per hour some of the time, at 50 kilometers per hour at other times, or he may have stopped for a while. On the *average*, however, he traveled at 65 kilometers per hour.

We see that the *average velocity* is the *time rate of the change in distance*. The actual velocity of the driver at a particular instant is called the *instantaneous velocity*. We can get a good idea of the instantaneous velocity of a driver by finding his average velocity over a very brief interval of time. Thus suppose we know that at 10:00 A.M. the driver is 50 kilometers away from a certain town and that 5 minutes later he is 55 kilometers away. Then he has traveled 5 kilometers in 5 minutes, or about 1 kilometer a minute; that is, 60 kilometers per hour. This is his *average velocity*, but it is also a very good approximation to his *instantaneous velocity*, since if he did stop during this time or was traveling much slower or faster than 60 kilometers per hour, it certainly couldn't have been for long.

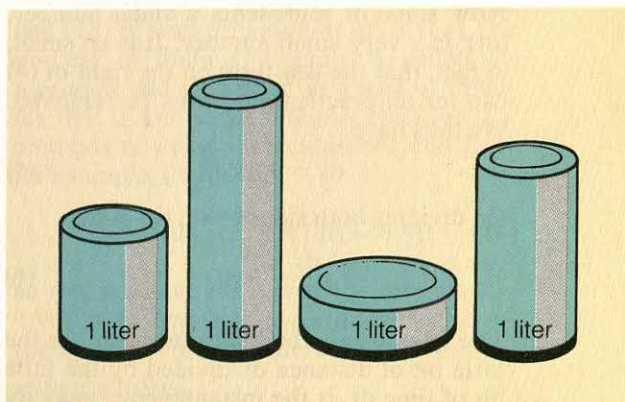
The average velocity is, as we have seen, the time rate of change in distance. The instantaneous velocity is the *instantaneous* time rate of change of distance. The problem of finding instantaneous time rates of change of distance, or of other quantities, is one of the most important parts of calculus and has many applications. We shall try to see how such instantaneous rates of change can be found.

PROCESS OF DIFFERENTIATION

Let us consider once again a ball dropped from a tall building. If we let s be the distance (in meters) that the ball falls in a time t (in seconds), then s will depend on t ; that is, s will be a function of t , or $s = f(t)$. A formula that reveals the relationship between s and t in case air resistance is negligible is this:

$$s = 4.9t^2 \quad (1)$$

From this formula we see that at $t = 0$ we



A manufacturer wants to make one-liter containers. Using calculus he can determine the minimal surface area needed to contain the required volume.

have $s = 0$, which is not surprising, since the ball falls zero distance in zero time. After 1 second—that is, $t = 1$ —we see from equation (1) that $s = 4.9$, which means that in 1 second the ball has fallen 4.9 meters. Similarly after 2 seconds—that is, $t = 2$ —we see that $s = 19.6$, so that in 2 seconds the ball has fallen 19.6 meters.

Suppose now we want to find the velocity of the ball at the time t —that is, the instantaneous velocity at time t . We increase the time t by a small amount, which we denote by dt . We can think of dt as a “little bit of t .”

We now try to find out the distance the ball will travel in the time $t + dt$ —that is, the time t plus the extra bit of time dt . This distance will be the original distance s plus an extra little bit of distance, which we call ds . Since the distance traveled in time $t + dt$ is $s + ds$, we see from formula (1) that

$$s + ds = 4.9(t + dt)^2 \quad (2)$$

which can be written

$$s + ds = 4.9t^2 + 9.8t(dt) + 4.9(dt)^2 \quad (3)$$

The extra little bit of distance ds which the ball travels is obtained by subtracting s from the left side of (3) and its equal value $4.9t^2$ from the right side of (3). In this way we find

$$ds = 9.8t(dt) + 4.9(dt)^2 \quad (4)$$

Now since dt represents a small number, $(dt)^2$ is a very small number. It is so small, in fact, that the last term on the right of (4) can for all practical purposes be removed. We thus have

$$ds = 9.8t(dt) \quad (5)$$

On dividing both sides by dt , we obtain

$$\frac{ds}{dt} = 9.8t \quad (6)$$

The quantity on the left of (6), which is the little bit of distance ds divided by the little bit of time dt , is the instantaneous velocity of the ball at time t . If we let this instantaneous velocity be v , we have

$$v = 9.8t \quad (7)$$

If we put $t = 3$ in (7) we find $v = 29.4$. This means that, after 3 seconds, the ball is traveling at 29.4 meters per second, which is its instantaneous velocity.

The process that we used to obtain the result (6) is known in the calculus as *differentiation*. That part of the calculus that deals with such processes is called *differential calculus*. The quantity ds/dt is called the *derivative of s with respect to t* , or simply the *derivative of s* .

If we like, we can use (7) to find the derivative of v with respect to t . To do this we use exactly the same procedure given above. We increase the time t by dt so that the total time is $t + dt$. Then the velocity v increases by a little bit, which we call dv , so that the new velocity is $v + dv$. From (7) we then see that

$$v + dv = 9.8(t + dt) \quad (8)$$

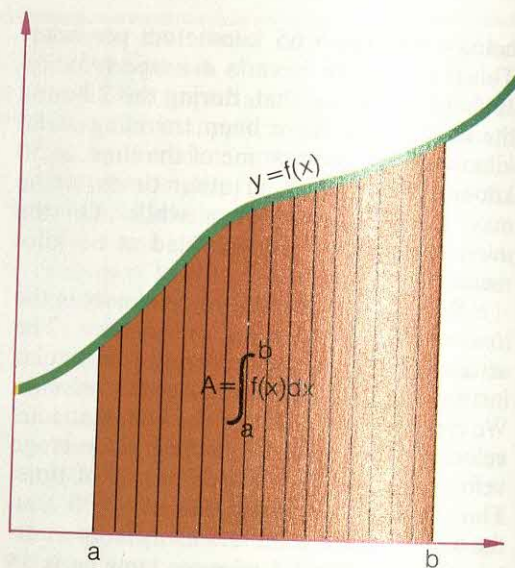
$$\text{or} \quad v + dv = 9.8t + 9.8dt \quad (9)$$

Subtracting v from the left side of (9) and its equal value $9.8t$ from the right side, we have:

$$dv = 9.8dt \quad (10)$$

$$\text{or} \quad \frac{dv}{dt} = 9.8 \quad (11)$$

The left side of (11) is the instantaneous time rate of change of the velocity, which is called the *instantaneous acceleration*.



What is the area under y from a to b ? We can divide the area into nearly rectangular regions whose areas can be approximated. The sum of these areas is very close to what we want. Integration gives the "best possible approximation," which is the actual area under the curve.

From (11) we see that the instantaneous acceleration is a constant. The significance is that when the ball falls toward the earth it increases its velocity by 9.8 meters per second in each second.

Since v is the derivative of s , we see that dv/dt is the derivative of the derivative of s , which is called the *second derivative of s* . It is written as:

$$\frac{d^2s}{dt^2} = 9.8 \quad (12)$$

APPLICATIONS

As we have already mentioned, the process of differentiation—that is, finding derivatives—is studied in that part of calculus known as *differential calculus*. There are many applications of differential calculus besides those involving velocity and acceleration.

One very important application of differential calculus is that of *maxima* and *minima*. To see what is involved in such applications, let us consider a particular problem. Let us suppose that we are in the business of making metal cans for a soup

company. We are asked by the company to make the cans cylindrical in shape, with the requirement that the cans have a capacity of one liter.

We could make the cans tall and thin, or short and fat. How shall we decide what to do? We know that the metal out of which the cans are made will cost money. It is therefore natural for us to ask ourselves whether we can make the required can by using the least amount—that is, the minimum amount—of metal for the total surface of the can. By the method of differential calculus we can determine the exact measurements, the diameter and height, which the can must have in order to contain a given volume and at the same time have the least possible surface area and therefore the least cost. Since we are interested in the least, or minimum, surface area, this is called a problem in finding *minima*.

As another illustration suppose a man has a rectangular piece of cardboard that measures 3 meters by 5 meters. He wishes to make an open box from it by cutting out equal squares from the corners and then bending up the sides. The question is: what size squares should be cut out so that the box will contain the greatest volume? This problem, which is one in finding *maxima*, can also be solved by differential calculus.

THE PROCESS OF INTEGRATION

We have already seen how, if we are given $s = 4.9t^2$, we can find $ds = 9.8t (dt)$. The process amounts to starting with the total distance s and finding the little bit of distance ds traveled in time dt . We gave the name differentiation to this process.

We now ask: if we are given the equation $ds = 9.8t(dt)$, can we get back to $s = 4.9t^2$? This is the reverse, or inverse, of differentiation. Since ds represents a little bit of distance, it is natural that to find the total distance s we must add up all the little bits of distance ds . The process of adding up, or summing, all the little bits of distance ds to produce s is represented mathematically by writing

$$\int ds = s \quad (13)$$

The symbol \int is called an *integral sign*, and (13) is read *the integral of ds is s* . The process of finding integrals is called *integration* and is studied in a part of calculus that is known as *integral calculus*. By methods of integral calculus, we can find, for example, that

$$s = \int ds = \int 9.8t(dt) = 4.9t^2 \quad (14)$$

so that we have recovered the formula $s = 4.9t^2$. It follows that integration is the reverse, or inverse, of differentiation.

APPLICATIONS

Just as there are many applications of differential calculus, there are also many applications of integral calculus.

One important application is that of finding areas bounded by complicated closed curves or of finding volumes bounded by complicated closed surfaces. Another application is that of finding the total length of a complicated curve or the total area of a complicated surface. The idea involved in such cases is that of summation, or addition, of little bits of area or little bits of volume and so involves integration.

We have seen how the process of integration enables us to go from the equation (6) back to the equation (1). The equation (6) is an equation that involves an instantaneous rate of change. It is a derivative. Equations involving derivatives of quantities that we want to determine are called *differential equations*. Such equations often arise in science and engineering because it is in many cases easier to arrive at a relationship between derivatives of quantities rather than between the quantities themselves. The process of solving differential equations is often also known as *integrating the differential equation*.

In the preceding paragraphs we have only been able to provide a glimpse into some of the many important ideas of calculus. In order to appreciate the power of the calculus in solving the many important problems of mathematics, science, and engineering, the student should consult some of the many books on calculus that are available.

SET THEORY

By Roy Dubisch

A herd of cows, a flock of birds, a school of fish—each of the words “herd,” “flock” and “school” could be replaced by the word “set.” A *set* is simply a collection of objects or ideas.

The concept of sets was developed into a new branch of mathematics in the late nineteenth century by a German mathematician, Georg Cantor. Since the beginning of the twentieth century, set theory has developed rapidly, and today it has important applications in nearly every branch of mathematics. In fact, most of our mathematics can be derived from set theory.

Two kinds of notations for sets are in common use. One is the enumeration notation, in which we write $\{1, 3\}$ for the set consisting of the numbers 1 and 3. $\{\text{Roberts, Roye}\}$ indicates the set consisting of the first two presidents of Liberia.

The second notation is the set-builder notation, in which we write $\{x|x \text{ is a whole number}\}$ for the set of whole numbers, and $\{y|y \text{ was one of the first two presidents of Liberia}\}$ for the set $\{\text{Roberts, Roye}\}$. We read $\{x|x \dots\}$ as “the set of all x such that $x \dots$.” Thus $\{x|x \text{ is a whole number}\}$ is read “the set of all x such that x is a whole number.”

Of the two notations, the set-builder notation is the most frequently used in mathematics. Note that the letter used in the set-builder notation is not significant. Thus, for example, $\{x|x \text{ is a whole number}\} = \{y|y \text{ is a whole number}\}$.

Capital letters are usually used as symbols for sets. We write, for example, $A = \{x|x \text{ is a whole number}\}$.

The *members*, or *elements*, of a set are simply those things that make up the set. Thus the members of the set $\{2, 3\}$ are the numbers 2 and 3. The members of the set $\{x|x \text{ is a United States citizen}\}$ are the citizens of the United States. When we say that two sets A and B are *equal* (written $A = B$), we mean that every member of A is a member of B and, conversely, that every

member of B is a member of A . For example, $\{5, 7\} = \{7, 5\}$. Note that the order of listing the members is not significant. If x is a member of a set A , we write $x \in A$. If x is not a member of A , we write $x \notin A$. Thus $2 \in \{2, 3\}$ but $4 \notin \{2, 3\}$.

SUBSETS OF SETS

Every member of the set $\{a, b\}$ is a member of the set $\{a, b, c\}$. We say that $\{a, b\}$ is a *subset* of $\{a, b, c\}$ and write $\{a, b\} \subseteq \{a, b, c\}$. Every member of the set $A = \{x|x \text{ is an even number}\}$ is a member of the set $B = \{x|x \text{ is a whole number}\}$. We say that A is a subset of B and write $A \subseteq B$.

Our examples illustrate the “natural” use of the word “subset” to indicate a part of a whole (as, for example, *subtotal* is a part of the total). It is useful, however, to consider that $\{a, b\} \subseteq \{a, b\}$. In general, every set is a subset of itself: $A \subseteq A$ for all sets A .

If A and B are sets, then $A \subseteq B$ means that whenever $x \in A$, then $x \in B$. If A is a subset of B but $A \neq B$, we write $A \subset B$ and say that A is a *proper* subset of B . Thus, for example, $\{a, b\}$ is a subset of $\{a, b\}$ but is not a proper subset of $\{a, b\}$, whereas $\{a, b\}$ is both a subset and a proper subset of $\{a, b, c\}$.

UNION OF SETS

Just as two numbers can be combined by addition or multiplication to yield a third number, so can two sets be combined in various ways to yield a third set. In particular, given two sets A and B we can form their *union*, $A \cup B$. This is the set whose members are members of A or B (or both). Thus, for example,

$$\{a, b\} \cup \{c, d\} = \{a, b, c, d\}$$

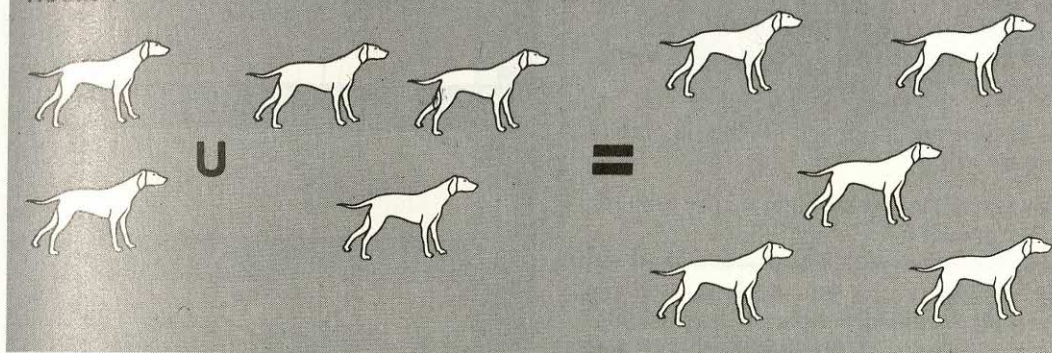
and

$$\{x|x \text{ is an odd number}\} \cup \{x|x \text{ is an even number}\}$$

$$= \{x|x \text{ is a whole number}\}.$$

In “set-theory language,” if A and B are sets, then $A \cup B = \{x|x \in A \text{ or } x \in B \text{ (or both)}\}$.

FIGURE 1



This concept of union is used in many parts of mathematics. In elementary school, for example, the idea of addition of whole numbers is made real to children by the use of set union. Thus the union of the two sets shown in Figure 1 corresponds to the addition fact $2 + 3 = 5$. In geometry the concept of union is used to define a triangle as the union of three line segments, each of which has a common end point with the other two, as shown in Figure 2.

Suppose we consider:

$$\{a, b\} \cup \{a, c\} = \{a, b, c\}.$$

Can we also write

$$\{a, b\} \cup \{a, c\} = \{a, a, b, c\}?$$

Yes, we can. It is true that $\{a, b, c\} = \{a, a, b, c\}$, but we do not "add" to the set $\{a, b\}$ by repeating the symbol "a". For example, consider the set $\{\text{you, reader of this article}\} = \{\text{you}\} = \{\text{reader of this article}\}$. You can't become two persons by naming yourself twice. So we agree, in listing the members of a set, not to list an element more than once.

INTERSECTION OF SETS

Similar to the concept of the union of two sets is the concept of the *intersection* of two sets.

If A and B are sets, then the intersection of A and B , $A \cap B$, is

$$A \cap B = \{x | x \in A \text{ and } x \in B\}.$$

Thus, for example,

$$\{a, b, c\} \cap \{a, e, f\} = \{a\}$$

and

$$\begin{aligned} &\{x | x \text{ is a brown-eyed human being}\} \\ &\cap \{x | x \text{ is a woman}\} \\ &= \{x | x \text{ is a brown-eyed woman}\}. \end{aligned}$$

Like the concept of union of sets, the concept of intersection of sets finds wide application in mathematics. The set of common divisors of 12 and 18, for example,

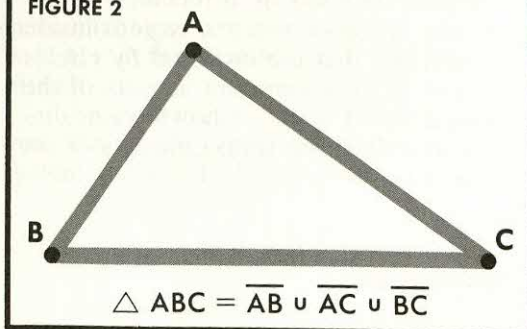
$$\begin{aligned} &\{x | x \text{ divides } 12\} \cap \{x \text{ divides } 18\} = \\ &\{1, 2, 3, 4, 6, 12\} \cap \{1, 2, 3, 6, 9, 18\} \\ &= \{1, 2, 3, 6\}. \end{aligned}$$

In geometry, we can symbolize the fact that two lines L_1 and L_2 intersect in a point p by

$$L_1 \cap L_2 = p.$$

What about $\{x | x \text{ is an even number}\} \cap \{x | x \text{ is an odd number}\}$? Since no number is both even and odd, it may seem as if no answer is possible. To handle this and many

FIGURE 2



other similar situations, however, we introduce the concept of the *empty set*. Also called the *null set*, the empty set is symbolized by $\{ \}$, or \emptyset . Thus

$$\{x|x \text{ is an even number}\} \cap \{x|x \text{ is an odd number}\} = \emptyset.$$

Likewise, if the lines L_1 and L_2 are parallel, $L_1 \cap L_2 = \emptyset$.

It is often useful to consider a fixed set, called the *universal set*, such that all the sets under consideration at a particular time are subsets of this universal set. For example, if we are discussing subsets of the whole numbers, such as the set of natural numbers $\{1, 2, 3, 4, \dots\}$ and the set of even numbers $\{0, 2, 4, \dots\}$, we could take our universal set as the set of whole numbers. Similarly, we would take our universal set as the set of all points in the plane if we were discussing subsets of the set of points in a plane, such as the sets of points forming circles and sets of points forming triangles.

COMPLEMENTS OF SETS

Suppose our universal set is the set of all students at Podunk High School, and A is the set of all female students at Podunk High. Then the set of all male students at Podunk High is called the *complement* of set A , and is indicated by A' . The complement A' of a set A consists of all elements in the universal set that are not members of A . In symbols, if $A \subseteq U$, then

$$A' = \{x|x \in U \text{ and } x \notin A\}.$$

Thus if our universal set is the set of real numbers and A is the set of rational numbers, then A' is the set of irrational numbers.

VENN DIAGRAMS

It is often helpful to picture sets and set relations by means of what are commonly called *Venn diagrams*. In Figure 3 the region enclosed by a rectangle represents the universal set, and regions inside the rectangle that are enclosed by circles (or other curves) represent subsets of the universal set. Figure 4 shows Venn diagrams that illustrate some concepts of set theory.

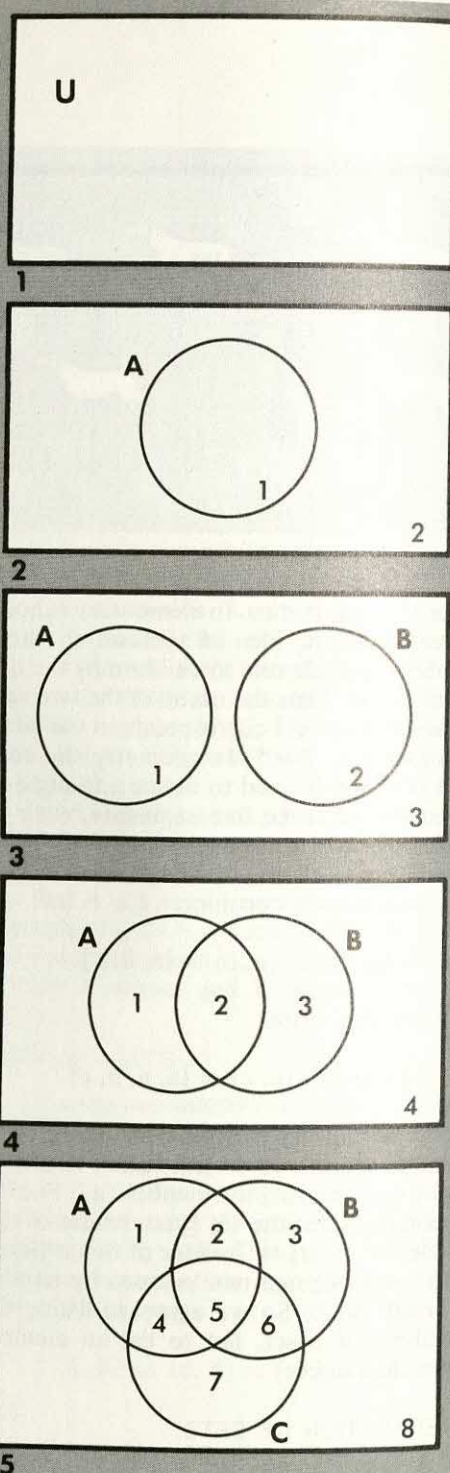


Figure 3. In these Venn diagrams, the rectangle represents the universal set, U . Subsets are represented by circles and are labeled A , B , and C . These subsets divide U into regions labeled with numbers. In diagram 3, sets A and B are "disjoint." In diagram 4 set A consists of regions 1 and 2; set B of regions 2 and 3. In diagram 5, set A consists of regions 1, 2, 4, and 5.

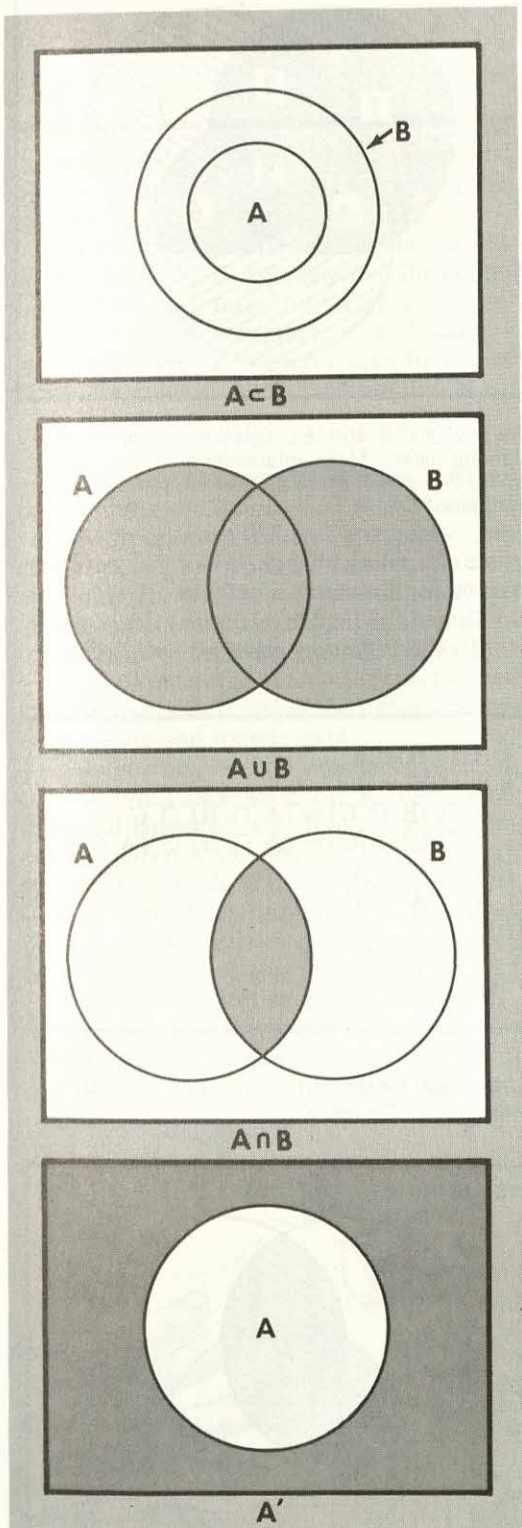
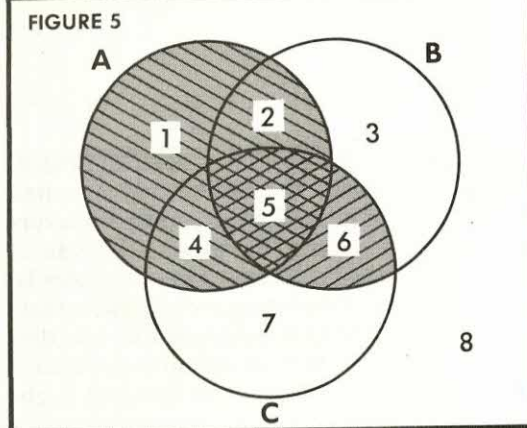


Figure 4. In the top diagram A is a proper subset of B . In the bottom three, the shaded areas represent the sets indicated below each diagram.



Set A consists of regions 1, 2, 4, and 5. Set $(B \cap C)$ consists of regions 5 and 6. What regions, then, does set $A \cup (B \cap C)$ consist of?

Venn diagrams can be used to make plausible—although not actually prove—various statements of equality between sets, such as the two equalities that, in algebra, are called the distributive properties:

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

and

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C).$$

(Compare $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$ in ordinary algebra.) Thus the shaded region in Figure 5 illustrates $A \cup (B \cap C)$ while the doubly shaded region in Figure 6 illustrates $(A \cup B) \cap (A \cup C)$. In both cases we obtain the same region.

The corresponding situations for $A \cap (B \cup C)$ and $(A \cap B) \cup (A \cap C)$ are shown in Figure 7.

ALGEBRA OF SETS

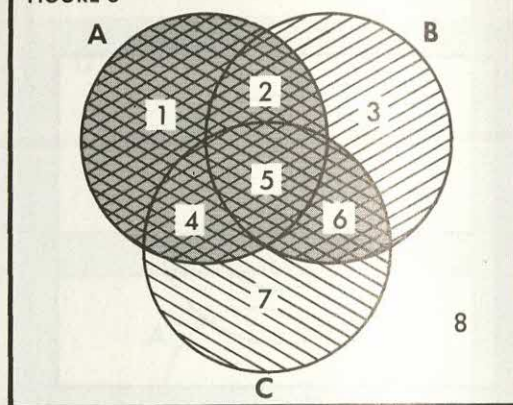
There are many analogies between the operations of union and intersection on sets and the operations of addition and multiplication on numbers. It is easy to see, for example, that just as $a + b = b + a$ and $a \cdot b = b \cdot a$ for all numbers a and b , it is also true that, for all sets A and B , $A \cup B = B \cup A$ and $A \cap B = B \cap A$. Likewise, just as 0 has the property that $a + 0 = a$ and $a \cdot 0 = 0$ for all numbers a , so does $A \cup \emptyset = A$ and $A \cap \emptyset = \emptyset$ for all sets A . Furthermore, just as $a \cdot 1 = a$ for all numbers a , so $A \cap U = A$ for all sets $A \subseteq U$ (the universal set).

The analogies are not complete, however. There are properties of our number system for which there are no analogous

properties for the algebra of sets, and there are properties of the algebra of sets for which there are no analogous properties for our number system. For example, given a nonzero number a , there exists a number b (the multiplicative inverse of a) such that $a \cdot b = 1$. Now U is analogous to 1 in the sense that $A \cap U = A$ for all sets A , and yet, unless $A = U$, there is no set B such that $A \cap B = U$.

On the other hand, the fact that $A \cup A = A \cap A = A$ for all sets A has no analog in arithmetic. We have stated that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ for all sets A , B , and C , and this does have an analog in $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$. We have also stated, however, that $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ for all sets A , B , and C , and this corresponds to $a + (b \cdot c) = (a + b) \cdot$

FIGURE 6



The sets $(A \cup B)$ and $(A \cup C)$ are indicated above by slanting lines. Their intersection is the region where the lines crisscross. It is the same set as that in Figure 5.

$(a + c)$ for all numbers a , b , and c . The latter statement, however, is not true. (Try $a = 1$, $b = 2$, and $c = 3$, for example.)

In the following table we list a number

- | | |
|--|---|
| (1) $A \cup A = A$ | (1') $A \cap A = A$ |
| (2) $A \cup B = B \cup A$ | (2') $A \cap B = B \cap A$ |
| (3) $A \cup (B \cup C) = (A \cup B) \cup C$ | (3') $A \cap (B \cap C) = (A \cap B) \cap C$ |
| (4) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ | (4') $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ |
| (5) $A \cup \emptyset = A$ | (5') $A \cap U = A$ |
| (6) $A \cup A' = U$ | (6') $A \cap A' = \emptyset$ |
| (7) $U' = \emptyset$ | (7') $\emptyset' = U$ |
| (8) $A \cup U = U$ | (8') $A \cap \emptyset = \emptyset$ |
| | (9) $(A')' = A$ |

The shaded area of figure 7A represents the set $A \cap (B \cup C)$, while the shaded area of figure 7B represents the set $(A \cap B) \cup (A \cap C)$. Both sets are the same, confirming the identity.

FIGURE 7A

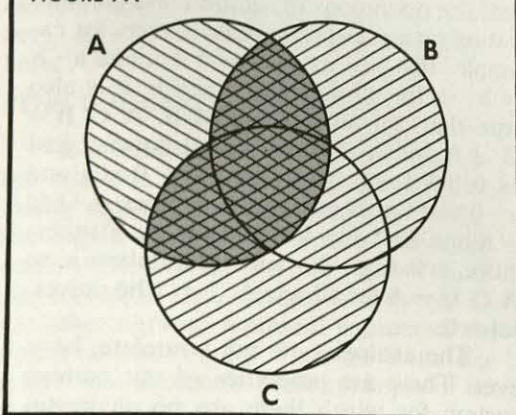
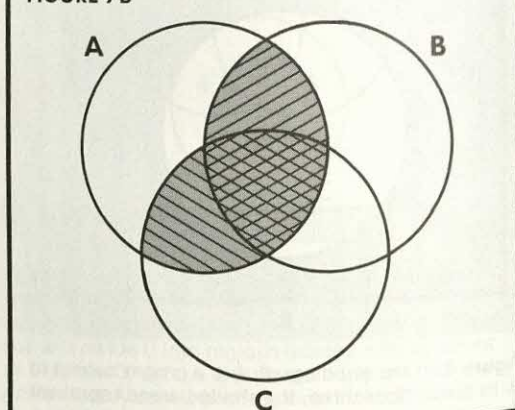


FIGURE 7B



of identities that hold for all sets A , B , and C that are subsets of some universal set U . Note the parallelism between the identities in the left and right hand columns.

All these identities are rather easily seen to hold as a consequence of the definitions of union, intersection, empty set, universal set, and complement, and can be made plausible, as we have seen by use of Venn diagrams.

INFINITE AND FINITE

What do we mean by "infinite"? Many people think of "infinity" as simply being a very large number. When a mathematician speaks of an infinite set, however, he does not think of a set with a large number of elements, such as the set of grains of sand on the seashore. Indeed, he finds it worthwhile to avoid entirely any direct reference to counting in describing the difference between finite and infinite sets.

To do this, he first considers the concept of *matching* sets, or, to use more-technical language, sets that are in *one-to-one* (1-1) *correspondence*. This is a very natural concept that is used even in the very early stages of instruction in mathematics when a child learns to pair off each block on a table with a finger of his hand.

A formal definition is as follows:

A 1-1 correspondence between two sets A and B is a pairing of the elements of A with the elements of B such that each

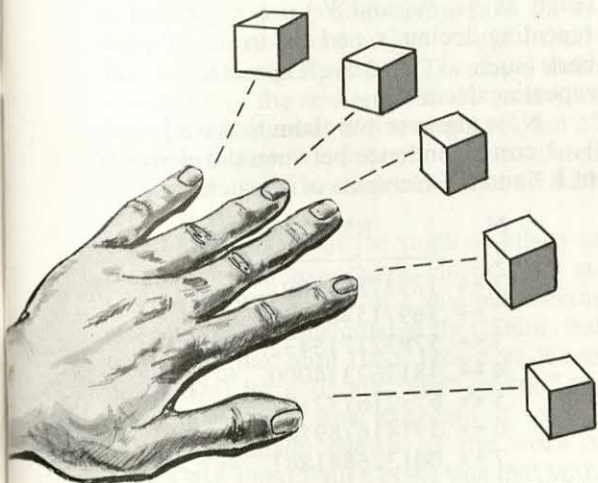


FIGURE 8

element of A is paired with precisely one element of B and each element of B is paired with precisely one element of A .

If such a 1-1 correspondence exists between the two nonempty sets A and B , we say that the two sets are *equivalent* and write $A \sim B$. (We also say that $\emptyset \sim \emptyset$, although we will not use this fact here.)

Now for finite sets we normally do not bother to attempt a pairing to see if the two sets are equivalent: we simply count the elements in each set. Thus we observe that each of the two sets shown in Figure 8 has five elements, so that we "know" that they are equivalent. Actually, however, the concept of 1-1 correspondence precedes the idea of counting. "Five," for example, is simply the number name that we attach to sets that are in 1-1 correspondence with the fingers of one hand. Primitive man, long before he knew how to count, could keep track of the size of his flock by matching each animal with a pebble. If he dropped a pebble into a container as each animal left the enclosure in the morning, and then removed a pebble as each animal returned in the evening, he would know whether or not the same number of animals returned as had left, without counting.

Now if we have any finite set A we sense intuitively that no proper subset of A is in 1-1 correspondence with A . Try, for example, matching the set of fingers minus the thumb on one hand with the entire set of fingers on the other hand. But what about infinite sets? Consider the set $N = \{1, 2, 3, \dots\}$ of natural numbers and the set $E = \{2, 4, 6, \dots\}$ of even natural numbers. Clearly E is a proper subset of N , and yet $E \sim N$ as shown by the 1-1 correspondence in Figure 9.

We generalize these two examples to make a formal definition of a finite set and an infinite set:

A *finite* set is a set that has no proper subset equivalent to itself. An *infinite* set is a set that has at least one proper subset equivalent to itself.

CARDINAL NUMBER OF AN INFINITE SET

If finite sets are equivalent, they have the same number of elements. That is, they

are *equinumerous*. What about infinite sets? For example, is the set E of even numbers equinumerous with the set N of natural numbers? From one point of view it is certainly natural to argue that E and N are not equinumerous, since we take something away from the set $N = \{1, 2, 3, \dots\}$ to get the set $E = \{2, 4, 6, \dots\}$. If we take something away, say 1, from the set $\{1, 2, 3\}$, the set we obtain $\{2, 3\}$, is not equinumerous with the set $\{1, 2, 3\}$.

Nevertheless, it turns out to be useful to agree that E and N are equinumerous and, in general, to make the following formal definition:

Two sets A and B (finite or infinite) are said to be *equinumerous* or to have the same *cardinal number* if and only if $A \sim B$.

Thus we conclude from this definition that not only do the finite sets of Figure 8 have the same cardinal number but also that the sets E and N of Figure 9 have the same cardinal number.

Do all infinite sets have the same cardinal number? Consider, for example, the set Q of all positive rational numbers—the set of all natural numbers together with all the positive fractions: $\frac{1}{2}$, $\frac{17}{19}$, $\frac{1}{8}$, $\frac{1}{252}$, 0.12 , and so on. Are Q and N equinumerous? Again, we may feel intuitively that the answer should be “no,” since Q contains many (indeed, an infinite quantity of) numbers not in set N . We can, however, show that $Q \sim N$, so that, according to our definition, Q and N are equinumerous.

A simple argument that $Q \sim N$ can be based on the diagram shown in Figure 10. It shows the positive rational numbers in an array extending indefinitely to the right and down. (Ignore the lines for the moment.)

Now suppose we are given a number $n \in N$ and wish to describe a rule for associating with n a definite rational number $r \in Q$. We simply follow the line in the diagram.

As we follow this line we count $1, 2, 3, \dots, n$ as we go through each fraction—except that we don’t count repetitions. For example, suppose we have $n = 5$. Then, along the line, we find $\frac{1}{1}$, $\frac{2}{1}$, $\frac{1}{2}$, $\frac{1}{3}$, and $\frac{3}{1}$ (omitting $\frac{2}{2} = \frac{1}{1}$) and conclude that corresponding to $5 \in N$ we have $\frac{3}{1} \in Q$. Similarly,

if $n = 13$ we obtain $\frac{1}{1}$, $\frac{2}{1}$, $\frac{1}{2}$, $\frac{1}{3}$, $\frac{3}{1}$, $\frac{4}{1}$, $\frac{3}{2}$, $\frac{2}{3}$, $\frac{1}{4}$, $\frac{5}{1}$, $\frac{6}{1}$, and $\frac{5}{2}$ (omitting $\frac{2}{2} = \frac{1}{1}$, $\frac{4}{2} = \frac{1}{2}$, $\frac{3}{3} = \frac{1}{1}$, and $\frac{4}{4} = \frac{1}{1}$). Thus to $13 \in N$ corresponds $\frac{5}{2} \in Q$.

Conversely, given any $r \in Q$, we take its representation as a fraction in lowest terms (such as $\frac{1}{2}$ for $\frac{3}{6}$) and count “how far” it is from $\frac{1}{1}$ along the line (again omitting repetitions). Thus if we have given $\frac{4}{6} \in Q$ we observe that $\frac{4}{6} = \frac{2}{3}$ and count $\frac{1}{1}$, $\frac{2}{1}$, $\frac{1}{2}$, $\frac{1}{3}$, $\frac{3}{1}$, $\frac{4}{1}$, $\frac{3}{2}$, $\frac{2}{3}$; so that corresponding to $\frac{2}{3} \in Q$ we have $8 \in N$.

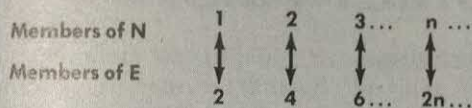
At this point you may suppose that all infinite sets have the same cardinal number as N . Certainly this is a plausible conjecture at this point. It turns out, however, that if we add to the set Q of positive rational numbers all the positive *irrational* numbers, such as $\sqrt{2}$, $\sqrt[3]{2}$, π , or $\sqrt[3]{1 + \sqrt{2}}$ and so on, we obtain a set with a cardinal number different from N . Even if we consider only the set R^* of rational and irrational numbers between 0 and 1, we obtain, as Georg Cantor showed in 1874, a “larger” set which cannot be matched with N . The union of the set of rational numbers and the set of irrational numbers is the set of real numbers.

The proof that R^* is not equinumerous with N rests upon the fact that every real number between 0 and 1 can be written as an infinite decimal. For example, $\frac{1}{2} = .5000\dots$, $\frac{1}{3} = .3333\dots$, $\frac{1}{7} = .142857142857\dots$, $\sqrt{2} = .7071\dots$, where the rational numbers (such as $\frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{7}$) are expressed as repeating decimals, and the irrational numbers (such as $\sqrt{2}$) are expressed as non-repeating decimals.

Now suppose we claim that we have a 1–1 correspondence between the elements of R^* and the elements of N , such as:

| N | R^* |
|--------------|--|
| \downarrow | |
| 1 | $\leftrightarrow .183478412001\dots = r_1$ |
| 2 | $\leftrightarrow .369715400000\dots = r_2$ |
| 3 | $\leftrightarrow .579321715432\dots = r_3$ |
| 4 | $\leftrightarrow .481762314000\dots = r_4$ |
| 5 | $\leftrightarrow .673216732167\dots = r_5$ |
| 6 | $\leftrightarrow .591416789143\dots = r_6$ |
| 7 | $\leftrightarrow .001326841841\dots = r_7$ |
| \vdots | $\leftrightarrow \dots\dots\dots = \dots$ |

FIGURE 9



The digits of r_1 , r_2 , and so on are chosen arbitrarily so that $r_1 \neq r_2 \neq \dots$

Now we form another number in R^* as follows: the first digit to the right of the decimal point is chosen as any digit different from the first digit of r_1 . The second digit is chosen as any digit different from the second digit of r_2 , and so on. Thus the resulting number could be .2783254... or .3723689.... None of these numbers so chosen, however, can be in our list, since .2783254... $\neq r_1$ because it differs from r_1 in at least the first decimal place; .2783254... $\neq r_2$ because it differs from r_2 in at least the second decimal place; and so on. Hence .2783254... was not present in the (supposedly complete) list, and so our correspondence is not 1-1 as alleged. No matter what listing we propose, the same process will show the listing is incomplete, and we conclude that no 1-1 correspondence between R^* and N is possible.

Mathematicians denote the cardinal number of N by \aleph_0 (read "aleph-null"—aleph is the first letter of the Hebrew alphabet). They denote the cardinal number of the real numbers by c. A famous question in mathematics is: does there exist an infinite set of numbers whose cardinal number lies between \aleph_0 and c? That is, is there a subset S of the real numbers R^* such that the cardinal number of S is neither \aleph_0 nor c? The *continuum hypothesis*, as formulated by Cantor, was the conjecture that no such set S exists.

The question of the truth or falsity of the continuum hypothesis claimed the attention of many first-rate mathematicians after it was first formulated by Cantor, but was not settled until 1963. However, it was not settled by a simple "yes" or "no" (as most mathematicians expected it to be). What was finally shown by the work of Kurt Gödel and Paul Cohen was that with- in the framework of the set theory accepted

by most mathematicians today, either the acceptance of the continuum hypothesis or its rejection yields equally valid systems of mathematics. This still leaves open the question of whether or not there exists another equally "useful" formulation of set theory in which the continuum hypothesis can be proved true or false. In a sense, then, the continuum hypothesis has been settled in only a relative fashion, and research still continues on this and related problems.

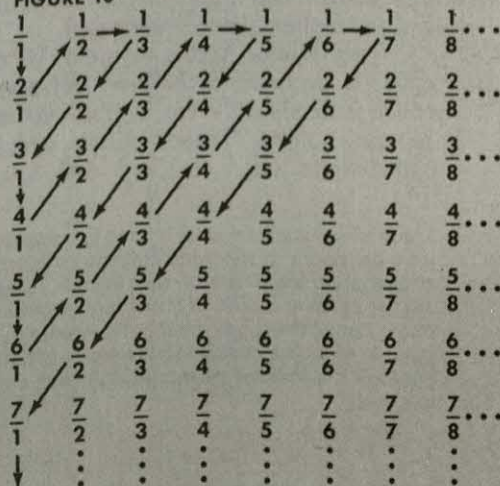
PARADOXES OF SET THEORY

There are many paradoxes associated with sets that still concern mathematicians today. One very famous one was formulated by the noted mathematician and philosopher Bertrand Russell (1872–1969). It is known as Russell's paradox. Imagine a barber in a village. The barber, said Russell, shaves only the men who do not shave themselves. The paradox is: Who shaves the barber?

If the barber does not shave himself, then he shaves himself. But if he shaves himself, he cannot shave himself. When stated mathematically, in terms of classes of sets, Russell's paradox challenged the very foundations of set theory.

The complete resolution of this and other paradoxes of set theory is still a matter of active concern today. In the meantime, however, most mathematicians continue to use the very useful concept of a set without waiting until all the foundations of set theory are made solid.

FIGURE 10



BINARY NUMERALS

by Irwin K. Feinstein

The binary numeral, or numeration, system offers an interesting glimpse into the world of mathematics. It is the simplest system in which addition and multiplication can be performed, for there are only four addition facts and four multiplication facts to learn. By contrast, in the decimal system there are 100 addition facts and 100 multiplication facts to learn.

The significance and power of the binary system in modern mathematics, however, lie in its practical application in computer technology. Most computers today are binary computers; that is, they use binary numerals in computing and in processing data. With binary numerals, modern computers can perform up to 1,000,000,000 additions each second.

Before going into the binary numeral system, we shall take a closer look at the Hindu-Arabic numeral system and the concept of place value in it. An understanding of place value is vital to an understanding of the binary numeral system.

HINDU-ARABIC NUMERAL SYSTEM

The numeration system in use throughout the civilized world today is called the Hindu-Arabic system, probably because it originated with the Hindus and was carried to the western world by the Arabs. It is the only numeration system most of us have ever known, and it fits most of our commercial and technical needs very well.

The Hindu-Arabic system is often called a *decimal*, or *base-ten*, system because it needs only ten symbols to represent any number. These symbols, which are called *digits*, are: 1, 2, 3, 4, 5, 6, 7, 8, 9, and

0. These ten symbols stand for the numbers one, two, three, four, five, six, seven, eight, nine, and zero respectively.

The Hindu-Arabic system is also called a *positional decimal system*, because the number each digit represents depends on its "position," or "place," in the numeral. The far-right place in any numeral is the unit, or one, position. The place value of the next position to the left is ten. In general the place values from right to left, in any decimal numeral, are: ones, tens, hundreds, thousands, and so on. For example, the place values represented by 4,081 are:

| 50281 | 1100000011100111001111011110011101110111011 |
|---------------------------|---|
| MARINER-C VIDEO TEST DATA | |
| MARINER MARS 1964 TV | |
| BINARY | |
| 50401 | 001111100001100001011110111110000001111100 |
| 50521 | 010001100010100000011111000001000010001000 |
| 50641 | 1100011000110001100011000110001000100010001100 |
| 50761 | 0100100100011100100100100100100100100100100100 |
| 50881 | 1100101001010011110101101010101010101010101010 |
| 51001 | 0001111001010010001000001000011101010100110 |
| 51121 | 110111101101101110111101111011110111111111111 |
| 51241 | 11100011000110101100011000111000110110110110 |
| 51361 | 1110101100011101011010111001110011010101110 |
| 51481 | 11110111101110111011001110011100111001110111 |
| 51601 | 11110011110111011100111001110011100111001111 |
| 51721 | 111110000100001000011000010000100001000010000 |
| 51841 | 00001000010000110000100010001000100010000100001 |
| 51961 | 0001100011000110001100011000110001100011000110001 |
| 52081 | 00100100100100100100100100100100100100100100100 |
| 52201 | 011010111101110111110000101111000011000 |
| 52321 | 100010111000111100111101111011110111011101101 |
| 52441 | 1111111111011111111111011110111101110111011101 |
| 52561 | 100011000110001100011010111010111010110101010 |
| 52681 | 101011000110001101011101011100110101011100 |
| 52801 | 11001110011100111001110111011101110111011101 |
| 52921 | 1101111011110000011111000001111100011111000111 |
| 53041 | 010100001000001000011000110000100000100001 |
| 53161 | 0101000010001100010100010001000100010001001011 |
| 53281 | 10110001100011001001001001001001001001001001001 |
| 53401 | 101001010001001001001001001001001001001001001011 |
| 53521 | 0100100001111111111111111000000101010100 |
| 53641 | 011110111011101110111011101110111011101110111 |
| 53761 | 011110110010100101010110001100011000110101011 |
| 53881 | 10110001100011010110101100011010110101100011 |
| 54001 | 1011100111011100110101110011010110101101011 |
| 54121 | 1011101111011101000001111011100111101111011 |
| 54241 | 010001010000010000010000010000010000100000100 |
| 54361 | 0100101111100000100011000100010001000000100 |
| 54481 | 010010001000100011000110001100011000110001100 |



NASA

The binary numeral system is used to transmit photos from space to receiving stations on earth. The photos are transmitted as radio signals, representing the zeros and ones of the binary system (upper photo right). The digits are then converted by computer to an image consisting of a series of dots. The photo of the Martian surface at right was obtained this way.

$$4,081 = (4 \times 1000) + (0 \times 100) + (8 \times 10) + (1 \times 1)$$

Another way of looking at a positional decimal system is through the idea of grouping. Suppose we have a set of 13 dots marked on a sheet of paper. We draw a ring around 10 of these dots. There will be 3 dots remaining. We have 1 set of 10 dots plus 3 remaining single dots. This fact may be expressed as $13 = (1 \times 10) + (3 \times 1)$.

Now suppose we have 37 dots, and that we "ring" them in groups of 10. We will have 3 sets of 10 dots, plus 7 single dots. We could write this as $37 = (3 \times 10) + (7 \times 1)$. Similarly, if we had 128 dots we could have 1 set of 1 hundred or 10 tens, 2 sets of 10, and 8 dots remaining. We could write $128 = (1 \times 100) + (2 \times 10) + (8 \times 1)$. All this emphasis on "tens" seems quite natural to us. After all, we do have ten fingers and ten toes. But how do you suppose man would be counting today and what sort of a numeration system would he be using if he did not have his anatomy structured as it is?

BINARY NUMERAL SYSTEM

Imagine all of us with one finger on each hand and one toe on each foot. Suppose further that all the numbers we use could be expressed with the "digits" 0 and 1. Since we are going to use only two symbols to write any number, we can call this a *binary*, or *base-two*, system. In a binary system the value of any place in a numeral is twice as large as the place to its right. Thus the place values—from right to left—in a binary system are: ones, twos, fours, eights, and so on.

The number 1 expressed in the binary system would be written as 1_{two} . Thus the number represented by 1_{ten} and 1_{two} is the same. The subscript *two* indicates that we are expressing numbers in the binary system; the subscript *ten* indicates the decimal system. How would we indicate 2_{ten} as a binary numeral? Let's try to discover the answer to this question.

Suppose we have a set of 2 stars. If we draw a ring around them, we have 1 set of 2 stars, and no single stars remaining. We

would write $2_{ten} = 10_{two}$, which is read as "1 two plus 0 ones." If we begin with a set of 3 stars and draw a ring around 2 of them, we would have 1 set of 2 stars plus 1 set of 1 star. This would be indicated as $3_{ten} = 11_{two}$, read as "1 two plus 1 one."

How would 4_{ten} be expressed? Notice that we have been making pairs of equivalent sets wherever possible. Thus if we had a set of 4 stars, we could first form 2 sets of 2 stars each. Now draw a ring around these sets. This approach suggests that 4_{ten} may be thought of as 1 four, no twos, and no ones, and written as 100_{two} . In similar fashion $5_{ten} = 101_{two}$, which is "1 four, no twos, and 1 one."

Note how 6 would then be treated. We would first have 3 sets of 2 stars in each set. Then we could pair up 2 of these sets, and wind up with 1 set of four, 1 set of two, and no ones. We write 6_{ten} as 110_{two} .

Using the same development, it is clear that $7_{ten} = 111_{two}$, interpreted as 1 set of four, 1 set of two, and 1 one. We can analyze 8_{ten} in the same way. First we have 4 sets of 2 stars in each set. Then we have 2 sets with (2×2) or 4 stars in each. Finally, we have 1 set with $(2 \times 2 \times 2)$ or 8 stars in it. This would be written as 1000_{two} and

TABLE I
Place-Value Chart

| Binary | | | | Decimal | |
|-------------|------------|-----------|-----------|------------|-----------|
| 8 eights | 4 fours | 2 twos | 1 ones | 10 tens | 1 ones |
| | | | 0 | | 0 |
| | | | 1 | | 1 |
| | | 1 | 0 | | 2 |
| | | 1 | 1 | | 3 |
| | 1 | 0 | 0 | | 4 |
| | 1 | 0 | 1 | | 5 |
| | 1 | 1 | 0 | | 6 |
| | 1 | 1 | 1 | | 7 |
| 1 | 0 | 0 | 0 | | 8 |
| 1 | 0 | 0 | 1 | | 9 |
| 1 | 0 | 1 | 0 | 1 | 0 |
| 1 | 0 | 1 | 1 | 1 | 1 |
| 1 | 1 | 0 | 0 | 1 | 2 |
| 1 | 1 | 0 | 1 | 1 | 3 |
| 1 | 1 | 1 | 0 | 1 | 4 |
| 1 | 1 | 1 | 1 | 1 | 5 |

interpreted to mean 1 eight, 0 fours, 0 twos, and 0 ones.

We now summarize what we have learned so far about binary numeration in Table I, and see if we can express our ideas.

How would we express 106_{ten} as a base-two numeral? Remember: the place values in binary numeration are ones, twos, fours, eights, sixteens, thirty-twos, sixty-fours, one hundred twenty-eights, and so on. The largest place value contained in 106 is sixty-four. There is 1 sixty-four in 106, and $106 - 64 = 42$. There is 1 thirty-two in 42; $42 - 32 = 10$. There are 0 sixteens in 10. There is 1 eight in 10; $10 - 8 = 2$. There are 0 fours in 2. There is 1 two in 2; $2 - 2 = 0$. There are 0 ones in 0. We conclude that $106_{ten} = 1101010_{two}$.

How would 47_{ten} be expressed in the binary system? There is 1 thirty-two in 47; $47 - 32 = 15$. There are 0 sixteens in 15. There is 1 eight in 15; $15 - 8 = 7$. There is 1 four in 7; $7 - 4 = 3$. There is 1 two in 3; $3 - 2 = 1$. There is 1 one in 1. So $47_{ten} = 101111_{two}$. But this method is tedious. There is another process that avoids much of this labor.

To convert any decimal numeral into its binary form, divide it successively by 2. For example, if we divide 47_{ten} successively by 2, we obtain:

$$\begin{array}{r} 23 \\ 2 \overline{)47} \\ \underline{46} \\ 1 \end{array}$$

$$\begin{array}{r} 11 \\ 2 \overline{)23} \\ \underline{22} \\ 1 \end{array}$$

$$\begin{array}{r} 5 \\ 2 \overline{)11} \\ \underline{10} \\ 1 \end{array}$$

$$\begin{array}{r} 2 \\ 2 \overline{)5} \\ \underline{4} \\ 1 \end{array}$$

$$\begin{array}{r} 1 \\ 2 \overline{)2} \\ \underline{2} \\ 0 \\ \\ 0 \\ 2 \overline{)1} \\ \underline{0} \\ 1 \end{array}$$

The remainders, in inverse order, make up the binary equivalent of 47_{ten} . Reading from bottom to top, the remainders are 1, 0, 1, 1, 1, 1. So $47_{ten} = 101111_{two}$. This checks with the result we obtained using the longer process.

BINARY ADDITION

Suppose we wish to add 11_{two} and 110_{two} . We expect the sum to represent the same natural number as if the numbers were expressed in base ten. As in base ten, addition in base two is always possible; the sum of any two numbers is unique; addition is commutative and associative; and 0 names the identity element.

With these structural properties in mind, it is relatively easy to develop addition in base two. (In the examples that follow, we shall at times discard the subscript "two" to avoid cluttering up our notation.) We begin:

$$1 + 0 = 0 + 1 = 1$$

and

$$0 + 0 = 0$$

The only remaining fact we need is $1 + 1$. But we have already found out that $1 + 1 = 10$ in the binary system. We therefore write:

$$1 + 1 = 10$$

We can summarize our addition facts in the form of a matrix for easy reference.

| | | |
|---|---|----|
| + | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 10 |

Now let's add $100 + 10$. We have:

$$\begin{array}{r} 100 \\ + 10 \\ \hline 110 \end{array}$$

Now try $10 + 11$:

$$\begin{array}{r} 10 \\ +11 \\ \hline 101 \end{array}$$

Discussion: $0 + 1 = 1$, so we write 1 in the ones' place in the sum. Then $1 + 1 = 10$. Write 0 in the twos' place in the sum, and 1 in the fours' place.

Now we shall do two more-difficult examples, using only the conventional addition algorithm.

$$\begin{array}{l} \text{(a) } 11010 + 10111 = \square \\ \text{(b) } 110100 + 1111001 = \square \end{array}$$

Solution (a):

$$\begin{array}{r} 11010 \\ +10111 \\ \hline 110001 \end{array}$$

Discussion: $0 + 1 = 1$; write 1 in the ones' place in the sum. $1 + 1 = 10$; write 0 in the twos' place and remember 1 four. $1 + 0 + 1 = 10$; write 0 in the fours' place and remember 1 eight. $1 + 1 = 10$; write 0 in the eights' place and remember 1 sixteen. $1 + 1 + 1 = 11$; write 1 in the sixteens' place and 1 in the thirty-twos' place. Solution (b):

$$\begin{array}{r} 110100 \\ +1111001 \\ \hline 10101101 \end{array}$$

Discussion: $0 + 1 = 1$; write 1 in the sum. $0 + 0 = 0$; write 0 in the sum. $1 + 0 = 1$; write 1 in the sum. $0 + 1 = 1$; write 1 in the sum. $1 + 1 = 10$; write 0 in the sum and remember 1. $1 + 1 + 1 = 11$; write 1 in the sum and remember 1. $1 + 1 = 10$; write 0 and then 1 in the sum.

Here are two more examples for you. See if your sums agree with these.

$$\begin{array}{r} 110101 \\ +1100011 \\ \hline 10011000 \end{array} \quad \begin{array}{r} 1101 \\ +1001 \\ \hline 10110 \end{array}$$

BINARY SUBTRACTION

Subtraction in the binary system is treated essentially the same as in the decimal system. What is needed is some additional flexibility with binary notation.

From 110 let us subtract 101.

$$\begin{array}{r} 110 \\ -101 \\ \hline 1 \end{array}$$

Discussion: We cannot subtract 1 one from 0 ones, so we take 1 unit from the next base position (twos) and think of it as '10' ones. 1 one from '10' ones leaves 1 one; write 1 in the ones' place in the difference. 0 twos subtracted from 0 twos leaves 0 twos; 1 four subtracted from 1 four leave 0 fours.

Here is another example:

$$\begin{array}{r} 1001 - 11 = \square \\ \begin{array}{r} 1001 \\ - 11 \\ \hline 110 \end{array} \end{array}$$

Discussion: 1 one from 1 one leaves 0 ones; write 0 in the ones' place in the difference. We can't subtract 1 two from 0 twos. The base position to the left of twos is fours, but there are no fours either. So we go to the eights' place. We change 1 eight to '10' fours. We take one unit from the '10' fours, which leaves 1 four, and change that to '10' twos. Now 1 two from '10' twos leaves 1 two; write 1 in the twos' place in the difference. We still have 1 four remaining, which we show by writing 1 in the fours' place in the difference.

Let us do a more difficult example:

$$\begin{array}{r} 10110 - 111 = \square \\ \begin{array}{r} 10110 \\ - 111 \\ \hline 1111 \end{array} \end{array}$$

Discussion: 1 is greater than 0. Change 1 two into '10' ones. 1 from 10 is 1. Change 1 four into '10' twos. 1 from 10 is 1. Change the 1 sixteen into '10' eights. Take 1 of these eights, which leaves 1 eight, and change it to '10' fours. Now we complete the subtraction. 1 from 10 is 1. 0 from 1 is 1. The difference is 1111.

Here are two more examples:

$$\begin{array}{r} 10011100 \\ - 110101 \\ \hline 1100111 \end{array} \quad \begin{array}{r} 10101101 \\ - 1111001 \\ \hline 110100 \end{array}$$

BINARY MULTIPLICATION

Suppose we wish to multiply 10 and 11. We expect the product to represent the same number as if the multiplication were performed in base ten. As in base ten, multiplication in base two is always possible, the product of any two numbers is unique, multiplication is commutative and associative; and 1_{two} names the identity element. Also we assume that the distributive laws hold. That is, multiplication is distributive over addition.

Keeping these assumptions in mind, let us develop multiplication in the binary system. We have:

$$1 \times 0 = 0 \times 1 = 0$$

and

$$1 \times 1 = 1$$

Our multiplication matrix would then be quite simple.

$$\begin{array}{r|rr} \times & 0 & 1 \\ \hline 0 & 0 & 0 \\ 1 & 0 & 1 \end{array}$$

Let's multiply: 11×10 .

$$\begin{array}{r} 10 \\ \times 11 \\ \hline 10 \\ 10 \\ \hline 110 \end{array}$$

Discussion: $1 \times 0 = 0$; write 0 in the product. $1 \times 1 = 1$; write 1 to the left of 0 in the first partial product. $1 \times 0 = 0$; write 0 under 1, in the second partial product. $1 \times 1 = 1$; write 1 to the left of the 0 just written. Now add.

Now we are ready to try a more difficult multiplication example.

$$101 \times 1101 = \square$$

Solution:

$$\begin{array}{r} 1101 \\ \times 101 \\ \hline 1101 \\ 0000 \\ 1101 \\ \hline 1000001 \end{array}$$

Discussion: For the first partial product: $1 \times 1 = 1$, $1 \times 0 = 0$, $1 \times 1 = 1$ and $1 \times 1 = 1$. Every entry in the second partial product is zero. The entries in the third partial product are the same as in the first. The addition is involved but not overly difficult.

Here are a few more examples for you. See if your products agree with these.

$$\begin{array}{r} 10111 \\ \times 1101 \\ \hline 10111 \\ 10111 \\ 10111 \\ 10111 \\ \hline 100101011 \end{array} \quad \begin{array}{r} 11100 \\ \times 111 \\ \hline 11100 \\ 11100 \\ 11100 \\ \hline 11000100 \end{array}$$

BINARY DIVISION

To be able to divide with binary numerals, you must be very familiar with binary multiplication and subtraction. The first few examples that we shall work will be done by repeated subtraction. Then we shall resort to the more conventional algorithm.

Let's divide 11011_{two} by 11_{two} .

$$\begin{array}{r} 1001 \\ 11 \overline{) 11011} \\ \underline{11} \\ 0011 \\ \underline{11} \\ 0 \end{array}$$

Discussion: 11 is contained in 11 just 1 time. $1 \times 11 = 11$. Subtract and get 0 for the difference. Bring down 0. 11 is contained 0 times in 0; write 0 in the quotient. Bring down 1. 11 is contained 0 times in 1; write 0 in the quotient. Bring down 1. 11 is contained 1 time in 11. The remainder is 0.

Let's look at another example:

$$\begin{array}{r} 110 \\ 101 \overline{) 111110} \\ \underline{101} \\ 101 \\ \underline{101} \\ 00 \end{array}$$

Discussion: 101 is contained in 111 just 1 time. $1 \times 101 = 101$. Subtract and get 10. Bring down 1. 101 is contained in 101 also 1 time. $1 \times 101 = 101$. Subtract and

get 0. Bring down 0. 101 is contained in 0 zero times.

Here are two more examples for you. See if your quotients agree with these. You may check your work by multiplying the divisor by the quotient to get the dividend.

$$\begin{array}{r}
 1110 \\
 10011 \overline{) 100001010} \\
 \underline{10011} \\
 11100 \\
 \underline{10011} \\
 10011 \\
 \underline{10011} \\
 1101 \\
 1101 \overline{) 10101001} \\
 \underline{1101} \\
 10000 \\
 \underline{1101} \\
 1101 \\
 \underline{1101}
 \end{array}$$

BINARY FRACTIONS TO DECIMALS

Converting fractional numbers in binary notation to equivalent decimal numerals is not particularly difficult. The position to the right of the "binary point" has a value of $\frac{1}{2}$; the next position has a value of $\frac{1}{4}$; the next a value of $\frac{1}{8}$; and so on. In each position, the numerator is 1, and the denominator is a power of 2 that depends on the position: for example, the third place to the right of the binary point is $\frac{1}{2^3}$, or $\frac{1}{8}$. We consider now an example.

Example: $1.1011_{two} = ?_{ten}$

Solution:

$$\begin{aligned}
 1.1011 &= 1 \times 1 + 1 \times \frac{1}{2} + 0 \times \frac{1}{4} \\
 &\quad + 1 \times \frac{1}{8} + 1 \times \frac{1}{16} \\
 &= 1 + \frac{1}{2} + 0 + \frac{1}{8} + \frac{1}{16} \\
 &= 1\frac{1}{16} \text{ or } 1.6875_{ten}
 \end{aligned}$$

DECIMALS TO BINARY FORM

Any decimal fraction can be converted to its binary equivalent as follows:

First write the fraction down; then multiply it by 2 using base-ten multiplication facts. For example, if the decimal fraction is 0.721, we write:

$$\begin{array}{r}
 0.721 \\
 1.442
 \end{array}$$

Then remove the integer to the left of the decimal point and record it, say on the right:

$$\begin{array}{r}
 0.721 \\
 1.442 \quad 1 \\
 0.442
 \end{array}$$

Repeating, we have:

$$\begin{array}{r}
 0.884 \quad 0 \\
 0.884 \\
 1.768 \quad 1 \\
 0.768 \\
 1.536 \quad 1 \\
 0.536 \\
 1.072 \quad 1 \\
 0.072
 \end{array}$$

The process is repeated until the decimal fraction is reduced to zero, or until the desired number of places in the binary fraction is obtained. The binary equivalent is obtained by reading the recorded digits from top to bottom.

Thus $0.721_{ten} = 0.10111_{two}$.

BINARY NUMERALS TO OCTALS

A number expressed as a binary numeral is readily expressed as an equivalent in base eight. For a natural number the technique is simple. Since $2^3 = 8$, we begin with the smallest unit and group the "digits" in clusters of three "digits" each.

Example: Convert 10110111_{two} to the equivalent base-eight numeral.

Solution:

$$\begin{array}{cccc}
 (10) & (110) & (111) &_{two} = 267_{eight} \\
 \uparrow & \uparrow & \uparrow & \\
 2_{eight} & 6_{eight} & 7_{eight} &
 \end{array}$$

Example: Convert $.1101101_{two}$ to base eight.

Solution:

$$.(110) (110) (100)_{two} = .664_{eight}$$

SUMMARY

The binary numeral system offers an interesting glimpse into the world of mathematics. Addition and multiplication, at first glance, appear to be very simple operations in the binary system because there are so few facts to learn. A more careful study reveals that numbers of any sizable magni-

tude require an almost endless sequence of 0's and 1's for their representations. Operations very quickly become bogged down in these long and labored numerals. But computers, because of their tremendous speed, can utilize binary notation to great advantage. Many of the digital computers of earliest design employed decimal notation, but it was soon apparent that binary notation had many advantages over other numeration systems for computer circuit design.

There are many interesting phenomena that lend themselves to binary explanation.

If a question can be answered by "yes" or "no"; if a circuit is either "closed" or "open"; if a light bulb is either "on" or "off"; if a choice involves either "male" or "female"; if the issue is either "binary odd" or "binary even"—these are the kinds of situations to which mathematicians apply binary notation and find it particularly convenient.

The "new mathematics" programs that are currently taught in some elementary and secondary schools treat the subject of binary numeration in some detail.

NIM: AN INTERESTING GAME BASED ON BINARY NUMERALS

An interesting game for two persons, called Nim, is based on binary numeration. In this game the players take turns drawing chips from three stacks before them. A player may draw as many chips as he chooses from any stack in a single move. In his next move he may take chips from the same stack or any other stack as he wishes. The player who takes the last chip left from the three stacks is the loser.

If this game is played by an experienced player and a novice, the beginner will rarely win. The secret of the game, as an experienced player will know, is to select the proper number of chips from the correct stack so that your arrangement will be "binary even." This then forces the other player to be "binary odd."

Here is what is meant by "binary even" and "binary odd." Suppose the stacks of chips, designated as A, B, and C, contain 9, 11, and 15 chips respectively. Represent 9, 11, and 15 in binary form: $9 = 1001_{two}$, $11 = 1011_{two}$, and $15 = 1111_{two}$.

$$\begin{array}{l} A \rightarrow \begin{array}{|c|c|c|c|} \hline 1 & 0 & 0 & 1 \\ \hline \end{array} \\ B \rightarrow \begin{array}{|c|c|c|c|} \hline 1 & 0 & 1 & 1 \\ \hline \end{array} \\ C \rightarrow \begin{array}{|c|c|c|c|} \hline 1 & 1 & 1 & 1 \\ \hline \end{array} \end{array}$$

After this is done, focus your attention on

the columns of numbers instead of on the original row representations. If the sum of each column is 0 or 2, then your position is "binary even." If not, your position is "binary odd." When it is your turn to move, hope that the position you are in is "odd" because if it is even, any move you make will force you into an odd position and may cause you to lose—if your opponent knows the game.

Suppose you are faced with the situation depicted in the illustration: 9, 11 and 15 chips. Your position is clearly "odd," so you must make a move which will leave your position "even." With a little practice you will see that you should remove 13 of the chips from stack C. If you do, your position will look like this:

$$\begin{array}{l} A \rightarrow \begin{array}{|c|c|c|c|} \hline 1 & 0 & 0 & 1 \\ \hline \end{array} \\ B \rightarrow \begin{array}{|c|c|c|c|} \hline 1 & 0 & 1 & 1 \\ \hline \end{array} \\ C \rightarrow \begin{array}{|c|c|c|c|} \hline 1 & 1 & 1 & 0 \\ \hline \end{array} \end{array}$$

Quite clearly, the sum of every column is either 2 or 0. Or, in other words, you are binary even.

Once you reach a safe (even) position, no matter what your opponent does, you will win, provided you make an "even" move every time.

DATA PROCESSING

by Elias M. Awad

What are the most profitable brands of soup? Which are money-losers? How does the manager of a supermarket determine the answers to questions such as these?

First he must gather *data*—facts and figures—on soup sales in his store. He must find out how many cans of each brand of soup are sold during a given period of time; how much he pays for a can; how much the customer pays for the same can; and so on. This is called *originating*, or bringing out, data. It must be done carefully. The data should be checked to be certain that it is

accurate. Then the manager must put this data into an understandable, finished form, such as a chart or a graph. The data, called the *input*, is now in such shape that it can easily be handled, or *manipulated*, by anyone.

From this input, the manager wishes to draw meaningful information on the profitability of different soup brands. This information is the *output* of his manipulated data input. The process *originating-input-manipulating-output* is the cornerstone of all data processing.

Data processing by machine is vital in modern life. It has many applications, ranging (top row, left to right) from computerized analysis of a heart patient's condition to speedy computerized confirmation of airline reservations and to the familiar cash-register computation. Computers process bank records (lower left) and rapidly provide all types of business data (lower right).

UPI; IBM; NCR; Raymond Juschus, Chase Manhattan Bank; NCR





Recording the data is essential to data processing. Records can be stored on punched cards (top and center), magnetic tape (lower left), or magnetic disks (lower right).



The success of the Apollo series of moon landings depended on complex computerized data-processing systems. Above: a photo of part of the mission control room during the Apollo 16 flight.

The manager may manipulate the data input in several ways. Paper and pencil may be enough. Or he may need a calculating machine. If the input is very large, he may have a computerized system.

The output from his manipulation of the data may be a written statement describing the profitability of different soup

brands. Or the output may be in the form of a mathematical equation that relates profits to costs.

The procedure described above is an example of data processing. The manager (1) *originated* the data, (2) arranged it for *input*, (3) *manipulated* it and (4) produced an *output* of orderly information.

We use data processing all the time in our daily lives without even realizing it. For example, whenever we see, hear, smell, taste, or feel anything, data in the form of nerve impulses races to the brain. The brain manipulates this data to obtain the information we need to think or behave in a certain way. In this case, the information and the reactions it produces may be considered the output of the brain's activity.

In the broadest sense, data is anything that produces information. Information is data that has been refined, summarized, and arranged in a logical fashion.

There are many ways to communicate, or relay, data. Speech and writing are probably the most familiar ways. Shorthand, electrical pulses, and magnetic patterns are some other methods.

DATA PROCESSING BY MACHINE

For many centuries people have used speech, written language, and numbers to process data. Most of this data arises from human activities such as historical events, government and business dealings, social life, and various personal affairs.

Until about a century ago, speech and writing were sufficient to process such data. But these methods became too slow as population began to increase and institutions became larger and more complicated. Man then turned to machines for help.

Like human beings, many machines—typewriters, electronic calculators, cash registers, and certain parts of computers—make use of ordinary letters, words, and numbers. But other devices use a “language” of their own: coded holes punched in paper tape and cards, electric and electronic pulses, magnetic patterns on tapes or disks, and so on.

Beginning about 1870, data-processing machines began to be adopted on a wide

performance, but to calculate your pay. You receive a definite wage, based on an hourly rate of compensation. From this, the company must deduct certain employee expenses: taxes, pension, insurance, and social security. At the end of each week or every two weeks or monthly, you get a paycheck.

These *payroll operations* can be complex and difficult, especially in companies that employ hundreds or thousands of people. In the following description of a typical payroll operation, we assume a computer is not being used. That is, the work is done by hand or on ordinary, electromechanical office machines.

As you come in to work, you may sign a time sheet, indicating the time you report. Or you may punch a time clock. The exact time of your entrance is thus recorded. After the day's work, you sign out the time you leave or punch the clock again.

In this way you have *originated* data on daily attendance at your job. By the end of the week or of some other convenient period, your time record, along with that of other employees, is checked for accuracy and edited. It perhaps is entered on some other form of record. The time records, in other words, are worked into an understandable data input.

This input is *manipulated* for output. One of the most important outputs in this case is your paycheck. Manipulation of data input here, as in many other situations, takes place in four basic steps: (1) classifying, (2) sorting, (3) calculating and recording, and (4) summarizing.

Let us look at each of these steps. When input data—the time records of employees—is *classified*, it is broken up into different groups. For example, the time records often are grouped according to the departments in the company. Each employee in a given department—such as accounting, production, or sales—usually has a number.

Note an important fact here. The data containing an employee's name, number, department, wages, and deductions exists before the current data on his working time comes in. This older data, which is gener-

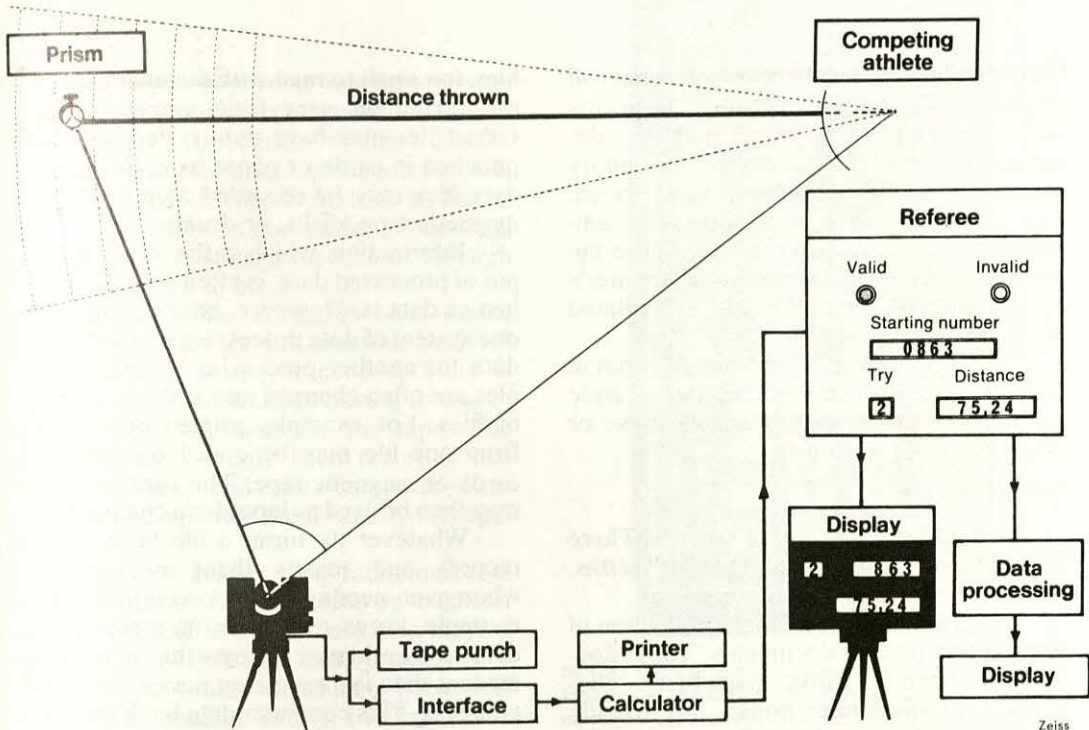


Electronic data processing is used at the Olympics. Seconds after an athlete throws a javelin, the distance thrown has been measured with the aid of a beam and the result posted. Hand-held beam reflector is shown at top of photo.

ally considered to be permanent, is compared or combined with the new data (working time). The result is the new output (current paycheck). Reworking of older and newer records is one of the most important features of data processing.

After classification the input is *sorted*. Within each group—in this case, a company department—data is put into some kind of order, or sequence. The employees' time records may be sorted in alphabetical order (by name) or in numerical order (by the employees' numbers).

Once the data has been classified and sorted, the payroll clerks *calculate and record* wages, hours worked, time off, and deductions for taxes, pension, insurance,



Measuring the field-event distances at the Olympics. After each throw, a prism marked the point of impact. A measuring instrument (lower left), placed at a known distance from the athlete, then emitted a beam that, when reflected, gave the distance between the prism and the measuring device. Instruments then computed the third side of the triangle—the distance thrown.

and social security. Another important phase of manipulation is the adding up, or *summarizing*, of all the data being processed. Much of the data connected with a payroll is very important to the managers of a firm. They use the figures to find out the labor costs of the company in a given period. Important decisions about company policies may then be made.

After the origination of data, its input and manipulation, we have an output. In our example, this includes, among other things, your paycheck. The check is usually in two attached parts. One part shows your net pay (pay after all deductions); this is the part you endorse and then cash at the bank or deposit to your own account. The other part of the check gives your gross pay, and may list the deductions—insurance, pension, taxes, and social security—that were taken out, resulting in your net pay. These deductions may vary greatly.

COMPUTER DATA PROCESSING

Basically, computerized data processing is much the same as that done by hand or by electromechanical methods. The main difference is that a computer handles all the work at one time, in one continuous operation at high speed. Certain steps, such as sorting, may be left out entirely. Very little is done by human beings.

Computer data processing takes place in three phases: *input*, the *processing* operation itself, and *output*. Data input is in such a form that the machine can “read” it for processing. The data may be recorded on punched cards or tape, on magnetic tape, or on magnetic disks or drums. Or it may be fed into the computer as typed or printed letters, words, numbers, or other symbols.

The computer then processes the input according to a program that has previously

been put into it. In the case of a payroll operation, the computer reads all the necessary data: employee's name, number, department, pay rate, number of hours worked, time off, deductions, and so on. The computer then calculates the employee's gross pay and net pay. These figures are then released from the computer's memory, or data-storage banks, and printed on a check.

In other computers, the output is not in printed form. Rather, it is recorded as code on punched cards, tape, magnetic tape, or magnetic disks or drums.

FILING OF DATA

A data file is a group of records. There are files on almost any subject: births, deaths, crime, and scientific research.

A familiar type of file is a collection of written and printed documents. These documents—papers, cards, pamphlets, clippings, and sometimes books—are usually kept in a definite order and stored in cabinets, in drawers, or on shelves.

Other kinds of files may be less familiar. Data may be photographed on micro-

film, too small to read with the unaided eye, but taking up very little storage space. Other files may have data coded as holes punched in cards or paper tape. Still other data files may be recorded as patterns on magnetic tape, disks, or drums.

Information, which is the finished output of processed data, is often stored in files just as data is. However, information from one system of data processing may serve as data for another processing system. Thus files are often changed into still other kinds of files. For example, printed documents from one file may be coded on punched cards or magnetic tape. The cards or tape may then be used as input for a computer.

Whatever its form, a file both stores records and makes them available to whomever needs them. A computer, for example, keeps records in its memory, or data-storage banks. It uses this more permanent data in processing newer data that comes in. This computer data bank is called *internal storage*. The storing of data outside the computer is called *external storage*. External data storage often exists in the form of magnetic tape, punched cards, and documents.

A file is classified in one of two ways:

1. A *functional file* is set up with a definite purpose in mind; the data is classified according to its function or use.

2. A *physically arranged file* is not set up for a specific function; its records are placed in any convenient order or location.

A functional file may be either a *master file* or a *transaction file*. A master file contains only related permanent records that need to be updated from time to time. In contrast, a transaction file holds only recent records, or transactions, which represent the changes to be made in updating a master file. A transaction file is temporary.

A file classified by physical arrangement keeps its records in either of two forms: sequential or random access.

Sequential. In a sequential file, the records are stored in a serial order, or sequence. A machine must scan all previous records before it reaches the one that is desired.

"Where do you feel pain?" A patient answers using an electronic "light pen." A computer then processes the patient's medical data and prints a summary for the doctor to study.



The most important medium for sequential data storage is magnetic tape. Suppose we have 10 records "written" on tape in the proper numerical sequence: 01, 02, 03, 04, 05, 06, 07, 08, 09 and 10. If, for instance, record 06 is needed for processing, the machine must scan records 01 through 05 before it gets to 06.

Sequential-file storage is ideal where regular updating of all records, such as customers' accounts, is needed. Time is not lost by the computer's having to search for individual records on the tape. The accounts have simply been recorded one after another, and the computer runs through and updates them rapidly, at regular periods.

Random access. In random-access files, records are usually stored on magnetic disks. The disk is similar to a long-playing record, except that each track on a disk is a circle. These are arranged in a pattern of smaller and smaller circles, one inside the other.

The records on a particular track, or section of track, have a specific code number. Given this code number by the person seeking a record, the machine can go directly to that track or section. Instead of going through all the records in the file, it need only "read" through those few having the same code number to find the desired record.

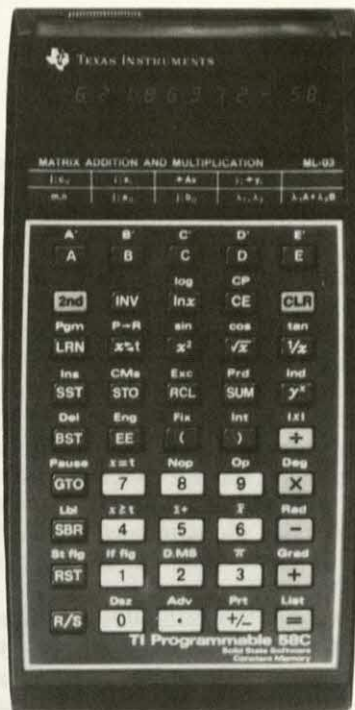
PROCESSING DATA FILES

Because many files are constantly tapped for information, they must be kept up to date. As new transactions take place, the data must be entered in the file. Depending on the file, this is done either by a person or by a machine.

If the records are stored on magnetic tape, disks, or drums, or on punched cards or tape, the data must be read by a machine such as a computer.

Sequential files are updated by an operation called batch processing. Random-access files are updated by on-line processing.

Batch processing. Records stored in sequence on magnetic tape or punched cards call for periodic processing. Records



Texas Instruments, Inc.

Not all data-processing machines are bulky. This electronic calculator and many others—some even smaller—can be put in a pocket. Such machines solve arithmetic problems, showing the answer in the window on top.

of transactions are gathered over a period of time. Then a batch of such records are all processed at the same time.

Batch processing is ideal for payrolls, commercial accounts, and other sequential transactions. There is an obvious saving of computer time if transactions can be handled this way.

A master file may also be batch-processed. If the master file is on magnetic tape, it can be put through a computer when it needs updating. Prior to the actual computer run, the batch of new data is sorted and sequenced in the same order as the master file. When the operation begins, the computer reads each piece of data and locates the corresponding record on the master tape. This operation continues until the tape is updated. A new, revised master tape is thus produced and is used for the next updating operation.

There are several advantages to batch processing. Because a large volume of records is processed at one time, there is a saving of computer time and of money. Also, the time and money needed to prepare

these records for processing are decreased. There is more effective control over processing errors. Totals of the figures can be taken both before and after each run, thus making sure that no transactions have been left out.

But batch processing also has disadvantages. It is difficult to process data at times other than those scheduled for batch processing. Updating a record right away may not be possible or may be expensive. Another disadvantage is that batch processing places a heavy load on a computer all at one time. Occasionally, serious problems may arise, such as delay or breakdown.

On-line processing. For random-access data files, an entirely different type of processing is used. On-line processing handles transactions as they come, regardless of their order. No sorting or batching is needed before the computer run.

Effective on-line processing requires the use of fast, so-called *direct-access* devices. These are magnetic disks or drums where every data-storage location and record can be reached directly.

In the on-line system a master record is updated shortly after a new transaction takes place. This system is ideal when transactions are not sequenced. It is especially useful when the situation demands more frequent runs than those made under batch processing.

Other advantages of on-line operation include faster supply of information to the user, because transactions are processed as they come. Further, a computer is not burdened with an excessive amount of work at one time, so that other data runs are not delayed. The generally faster on-line method makes the whole operation efficient from the standpoint of time saved.

But on-line processing has certain disadvantages. Large and expensive data-storage devices are needed. Processing costs can also be high, since usually a relatively few transactions are run through at one time. Also, it is hard to trace the flow of data from input to output through an on-line system. If the system breaks down, serious delays and costly errors result.



Sylvania

A policeman dials a coded message, which is transmitted by regular radio to a terminal where a computer processes the message.

PROCESSING OF AIRLINE RESERVATIONS

On-line processing is especially useful in the reservation of airplane passenger tickets. The actual time it takes to get a response to a request is only a matter of seconds.

American Airlines, for example, has SABER, a Semiautomatic Business Environment Research system, which allows hundreds of ticket agents to make reservations through a central computer complex almost instantly.

When a prospective passenger calls for one or more reservations on a flight, the ticket agent punches coded information through a terminal. The terminal is connected, along with other terminals, through low-speed lines to a terminal interchange located in the same area.

The interchange holds the agent's message until the central computer complex, in New York (the headquarters), is ready to take it. When this happens, the message goes through high-speed lines to the complex. The computer then begins to search the data filed on its magnetic disks for an answer to the ticket request.

When the answer is ready, it is transmitted from the computer complex to the agent who sent the original message. The agent, in turn, informs the prospective passenger whether there are seats on the desired flight and whether the reservations are confirmed. The entire operation from beginning to end takes about 3 to 5 seconds.

The program used in running SABER

consists of over 150,000 instructions. It handles, among others, the following key items of information: the passenger's name; the name of the person making the reservation; the phone number where the passenger may be reached; the name of the person picking up the tickets; special-menu requests; wheelchair requests; special shows booked at the flight destination; car-rental requests.

The complex structure of this system is shown by the fact that the entire inventory of such information can be updated, beginning as much as one year ahead of a current reservation date.

At present SABER terminals and other similar systems, such as the Apollo system, are operated by ticket agents all over the world. More than 65 percent of U.S. travel agencies use these computerized reservation tools. In 1983 they sold about two-thirds of all airline tickets issued in the United States.

DESIGN OF DATA PROCESSING

Any organization, no matter what its size, processes data. If the output is to have any meaning, the data must be processed in a systematic way.

A data system has two main jobs: creating data for files and keeping the files up to date. Thus, new output is created from time to time, such as paychecks in a payroll operation.

In order for an organization to get the desired output, a definite routine, or procedure, whether computerized or not, must be set up for the organization. This is the job of *systems analysts*. These experts work in teams and discuss the problems involved.

Systems analysis and design include the following activities: (1) collection of data—showing the data to be used; (2) analyzing the data—showing how the data is to be handled; (3) systems design—creating the routine to be followed in getting the necessary output.

The systems analyst is the key person in the systems concept. The analyst directs those people in charge of developing the needed procedures. The analyst must keep

the managers of an organization informed about the system of data processing that best serves their interests and about the state of the system at all times.

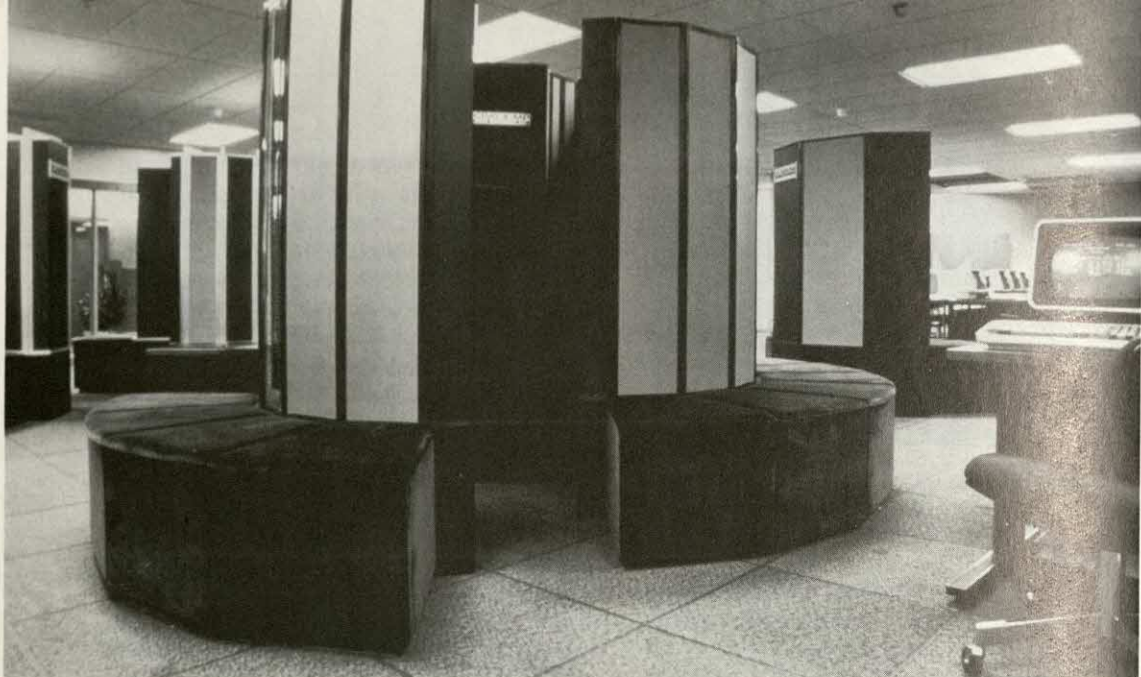
Furthermore, the analyst explores new methods of designing more efficient data-processing systems and keeps informed about the abilities and costs of different equipment being sold on the market. The analyst must justify the need for up-to-date reports, cut out any unnecessary procedures, and switch people from dull, routine work to more creative jobs.

To suit the standards and needs discussed above, data-processing equipment must be chosen carefully. The best system is one that produces the desired output in proper form at the rate of speed required, and at the lowest possible costs in time, labor, and money.

Also important are the size of the organization, the volume of data to be processed, the complexity of the operations, and the accuracy needed. All the considerations mentioned above must be weighed before a particular data-processing system is chosen.

The system chosen may be manual. This may be ideal for small organizations that handle limited amounts of simple data. Or the system may use mechanical or electromechanical equipment, such as punched-card or punched-tape processors, calculators, and conventional typewriters. A large organization handling tons of data needs a computerized system. The chief operation chosen may be batch processing or on-line processing.

All in all, the combined efforts of man and machine make it possible for organizations to handle data effectively. This kind of system benefits individuals, too. Workers in large companies receive paychecks on time because of complex data-processing machines. Mathematicians and scientists process data on computers, thus completing complex calculations in minutes instead of months. Banks use data-processing systems to determine the interest on savings accounts. In these and many other instances, people who handle figures are relieved of almost endless drudgery.



Cray Research, Inc

The cylindrical arrangement of this very powerful Cray computer enhances operating speed by permitting shorter than average connections among its elements.

COMPUTERS

by R. Clay Sprowls

"Good morning, Alice." As a computer terminal prints out this greeting, eight-year-old Alice begins her lesson in modern math. At the same time, one of her classmates begins a computer-assisted lesson in spelling.

Two thousand kilometers away, a poultry farmer in Oregon uses a computer to determine how much it will cost to double the size of his operation.

In New York City a police patrol car spots a suspicious vehicle. The police officer calls headquarters and relays the license plate number to a computer operator. Within seconds a reply comes back: the car isn't wanted.

In Mexico City a doctor must treat a seriously ill U.S. visitor. The physician telephones a computerized medical-history reporting service in New York. Within minutes the doctor learns that the patient is allergic to sulfa drugs, has type-B⁺ blood, and is a diabetic.

In Belgium a young man dies in an automobile accident. His tissues are typed and matched via computer with the tissue types of 300 patients awaiting kidney transplants. One of the young man's kidneys is flown to a hospital in Luxembourg, the other to England. The entire procedure takes less than 12 hours.

Today computers are used in teaching languages, building airplanes, keeping track of manufacturing orders, drawing maps and graphs, even hunting for hidden chambers in the ancient pyramids of Egypt. In short, they have become an integral part of our lives, whether we are scientists, businesspersons, students, or homemakers. Computers, in fact, have become the "giant brains" on which large parts of modern society depend.

The computer, however, is basically simple and should not be held in awe. It should not be kept at arm's length out of either fear or ignorance. Fundamentally,

There are two kinds of computers: digital and analog. The *analog computer* deals directly with measurable quantities like forces, voltages, pressures, temperatures, and other continuous variables. It is designed to be an analogy or a physical likeness of the problem it is to solve. Equations that describe the relations among the computer variables are assumed to be closely related to the problem variables. Changes in the computer variables are interpreted as the outcomes that would result from changes in the problem variables.

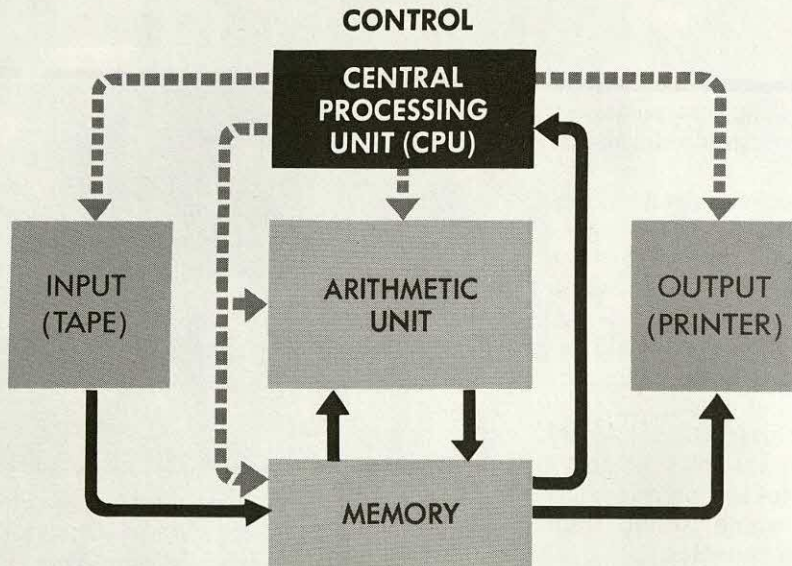
When people use the word "computer," however, they are usually referring to the digital type. Although the analog type still has some uses, these are limited. The emphasis in this article is on the digital computer.

The computers used in our society today are many and varied. Each of them has certain basic components that are always present. Beyond the basic parts, computers differ in the number and variety of units that are added to form a computing system. A knowledge of the basic functioning parts provides a frame of reference with which to approach the more complex computing systems. Probably the most important part of any computer is the *central processing unit*.

[illegible]

nothing more than addition, subtraction, multiplication, and division. The logical operations are very little more than an ability to compare two numbers and determine whether they are equal, and, if unequal, which one is the larger.

Associated very closely with the central processing unit is a storage unit. This is the primary storage unit for the computer and is often called the *memory*. This unit receives data, holds the data indefinitely without erasure or loss, and supplies the data upon command from the CPU for



The five basic parts of a computer. The solid lines indicate the direction of flow for the data values and the instructions. The dashed lines indicate the control the CPU has over other units.

processing. Together, the primary storage and the CPU are the heart of the computer system.

The input and output units are the mechanisms by which data and other information are transferred between the computer storage and the outside world. One common input medium is the magnetic disk. Data recorded on the disk are read into the computer storage. One very important output device is some form of printer, perhaps as simple as a typewriter or as complicated as a high-speed laser printer that will print thousands of lines per minute.

Finally, human control over the computer is exercised through the console. Lights, switches, and buttons not only enable the computer operator to control the computer, but also to monitor what the computer is doing at any moment. Very often the console is designed into the CPU so that physically the processor and console are in one box.

A simplified diagram of a computer is shown above. The diagram represents a small computer with magnetic tape for input and a printer for output. The solid lines indicate the direction of flow for data values and instructions. For example, input values are transferred from the tape to memory; output values from memory to the printer. Values are also transferred be-

tween memory and the arithmetic unit in both directions. Instructions are transferred from memory to the control unit. The dashed lines indicate the exercise of control. Thus, they connect the control unit with the input, output, and arithmetic units as well as with the memory.

Any computer has these basic functioning parts. Large and complex computer systems have both a greater variety of units than shown in the diagram and more of each unit. Large computers with multiple terminals usually costing in excess of \$100,000 are called mainframe computers. Medium-sized computers that can also service multiple users but can fit into an office situation and cost under \$100,000 are called minicomputers. Small computers serving a single user in home or business are called microcomputers, or personal computers.

STORING INFORMATION

The digital computer represents numbers in a binary form. One popular binary code is the Extended Binary Coded Decimal Interchange Code, abbreviated as EBCDIC and pronounced phonetically as *EEB-SEE-DIK*. This is a code in which eight binary digits, or eight *bits*, code one character. Each group of eight bits is called a *byte*. Each character that the computer can represent is coded into a unique byte.

Eight bits will provide 256 such unique codes, and these can be used to represent 256 different characters in computer storage. A selected list of EBCDIC characters and their codes is shown in Table 1. The memory unit stores all data in this or some other binary form.

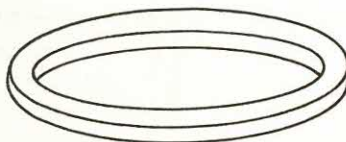
TABLE I
Selected EBCDIC Coded Characters

| | |
|-------|----------|
| blank | 01000000 |
| . | 01001011 |
| + | 01001110 |
| - | 01100000 |
| = | 01111110 |
| A | 11000001 |
| B | 11000010 |
| C | 11000011 |
| Z | 11101001 |
| D | 11110000 |
| 1 | 11110001 |
| 2 | 11110010 |
| 9 | 11111001 |

The memory unit of a computer should have five desirable qualities: (1) it should be reliable, in the sense that a character is not easily erased; (2) it should be large, so that it can store a lot of data; (3) it should be organizable into independent locations, so that characters will not be placed on top of each other; (4) it should be fast, so that access to values takes very little time; and (5) it should be inexpensive.

Not all of these desirable properties can be found in a single storage medium. One form of memory is composed of *magnetic cores*, which are reliable, fast, and organizable into independent locations. When core memories are large, they are also expensive, so that low cost comes only with limited size.

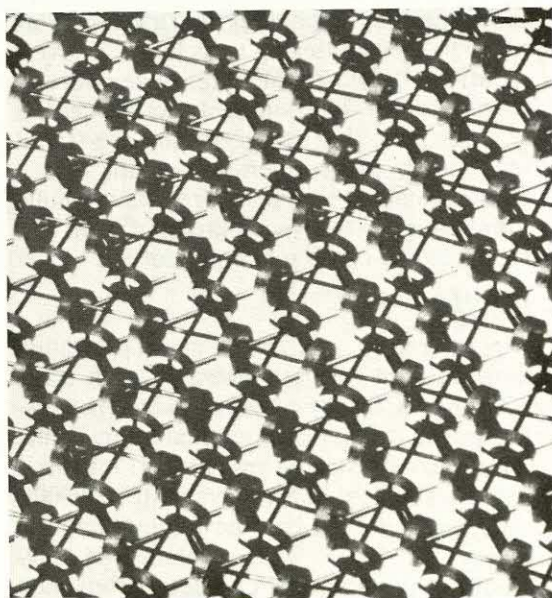
A magnetic core looks like a tiny doughnut that is as small as 0.5 millimeter in diameter with a 0.4-millimeter hole. (See Figure.) It is made of a ferrous material that can be magnetized, either in a clockwise or a counterclockwise direction. Because of this, the core can store information. The two magnetized "states" are unique and distinguishable and can represent the 0 and 1 values of a binary digit.



Although one core can store only binary-digit values of 0 or 1, a collection of cores can store a very large number of bits.

An electric current moving along a wire generates a magnetic field around the wire. When the direction of the current is reversed, the direction of the magnetic field is reversed. If a magnetic core is strung on a wire, the magnetic field generated by a current moving along the wire magnetizes the core in the same direction. If a whole string of cores is subject to the magnetic field generated by the current passing along their wire, each will be magnetized. For use in a computer memory, a single core must be magnetized instead of the whole string. This is accomplished by making use of the magnetic properties of the material of which the core is made.

Part of an actual core plane, greatly magnified. Each core can store one bit of information. Notice how each core is located at the intersection of two wires.



If the electric current passing through the wire is too weak, the core is not permanently magnetized. Only if the current is strong enough will it generate a magnetic field sufficient to magnetize the core permanently regardless of its former condition. A current that is too weak will leave the core unaffected. For this reason, each core is placed at the intersection of two wires. The wires are threaded through the holes of the cores. The strength of the current passing through the two wires is carefully regulated so that a current in one of them is not sufficient to magnetize a core. Only with two current pulses acting simultaneously will the current strength exceed the threshold needed to magnetize a core permanently. Current flowing along both wires will permanently magnetize just the single core that is located where the wires cross.

The cores are organized into planes. Each core in a plane is uniquely located by the intersection of two wires: one strung horizontally and one strung vertically in the plane. If eight such planes are stacked on top of each other, the vertical column of eight cores at the same row-column intersection of each plane represents the eight bits of one byte and therefore one character of binary-coded data.

Collections of bits (cores) organized into planes and stacks of planes form computer-memory capacities that run from as small as 4,000 bytes up to as large as several million bytes. These bytes have addresses to identify them uniquely to the central processor, so that characters may be stored in memory independently and not placed one upon another. The address is a unique number assigned to each group of eight bits (one vertical column of eight cores).

BINARY CHARACTERS

Internally, characters are represented as eight-digit binary numerals. Each character has a binary-numeric value that distinguishes it from another character. In the characters illustrated in Table 1, "blank" has the smallest numeric value (01000000), and the digit "9" has the largest numeric value (11111001). These values not only identify the individual characters, but they

permit comparisons of one with another. For example, the largest special character is the "=" sign (01111110). Numerically it is smaller than the first letter of the alphabet, A (11000001). Z has the largest value of any alphabetic character (11101001), and it is smaller than the smallest of any of the decimal digits (11110000).

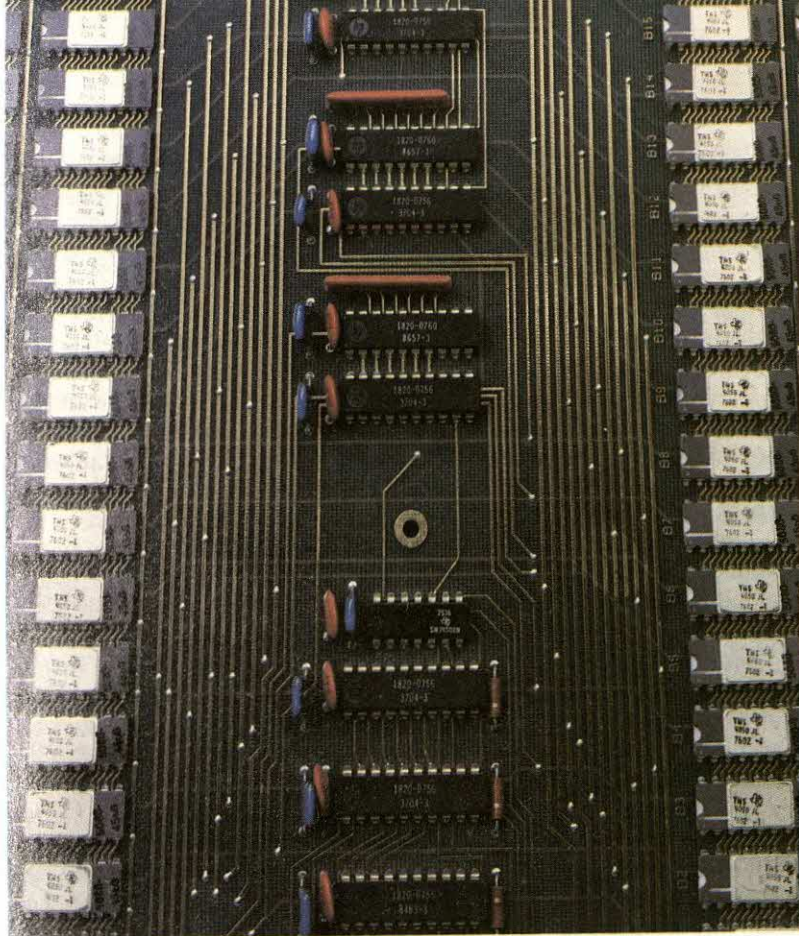
Comparing characters in the computer is unlike comparing them visually, where the shape of the character is so important in reading. The internal binary code is important in the computer. The order of such codes is established so that the binary-numeric values progress in an order that permits comparison, and therefore sorting, into a meaningful sequence. Values that are names of people, for example, are sorted by comparing the byte codes in increasing value to organize them from A to Z—that is, from 11000001 to 11101001.

In some applications and uses of the computer, a knowledge of the exact contents of the memory is necessary. By exact contents is meant the status of the individual bits of the data value. A printout of the contents of memory is called a *dump*. A dump of the memory in binary form is extremely difficult to interpret. Even to compare one eight-bit byte with another to see what each represents is difficult. To search through hundreds and possibly thousands of such binary values, as is sometimes necessary, is impossible. Therefore byte-organized computers are usually capable of representing bits in the notation of hexadecimal numerals.

HEXADECIMAL NUMERAL SYSTEM

The hexadecimal numeral system uses the base 16, as contrasted with the base 10 of the decimal system and base 2 of the binary system. The 16 different characters needed to represent each hexadecimal digit may be chosen rather arbitrarily, but current conventions use the 10 decimal digits 0 through 9 and the letters A, B, C, D, E, and F. The hexadecimal digits and their decimal counterparts are shown in Table II.

A single hexadecimal digit can uniquely represent four binary bits because there are 16 different patterns of four binary dig-



Hewlett Packard

The primary storage unit of a computer is called its memory. This unit receives data, holds data indefinitely, and supplies data for processing. Above is the complex memory board for a Hewlett Packard 3000 Series 2 computer.

its in the numbers from 0000 to 1111. These are also shown in Table II. A print-out of memory in hexadecimal notation is a representation of the exact contents of storage. One need only translate hexadecimal into binary to know exactly the bits that are stored. For example, suppose the computer is treating four bytes as a single unit of data. The internal value might be

1100010100100000000101101111101

in binary numerals. This number is more easily represented in hexadecimal by converting each group of four binary digits into one hexadecimal digit.

C 5 2 0 1 6 F D

Now a change of one bit of the binary number is not easily detected in the binary form. Can you find which bit is changed below?

1100010100100000000111101111101

A printout in hexadecimal is

C 5 2 0 1 E F D

TABLE II
HEXADECIMAL NUMERALS

| DECIMAL | HEXA- DECIMAL | BINARY |
|---------|------------------|--------|
| 0 | 0 | 0000 |
| 1 | 1 | 0001 |
| 2 | 2 | 0010 |
| 3 | 3 | 0011 |
| 4 | 4 | 0100 |
| 5 | 5 | 0101 |
| 6 | 6 | 0110 |
| 7 | 7 | 0111 |
| 8 | 8 | 1000 |
| 9 | 9 | 1001 |
| 10 | A | 1010 |
| 11 | B | 1011 |
| 12 | C | 1100 |
| 13 | D | 1101 |
| 14 | E | 1110 |
| 15 | F | 1111 |

The change from 6 (0110) to E (1110) shows up very readily.

Or suppose that a core dump produces the following hexadecimal number

8 0 1 1

One can easily write the binary form as

1000000000010001

and if this value is known to be a single binary integer with a sign bit and 15 data bits, the decimal integer value is -17.

Computers do not operate with hexadecimal numbers. These numerals are merely a convenient way to represent binary numerals. They are important in present-day computers with memories organized around bytes, or eight bits, because two hexadecimal digits represent the exact bit contents of one byte in memory.

Magnetic-core memories are fast. Computers today can deliver a byte from memory to the central processor in about one microsecond—one-millionth of a second. This "delivery time" is called the *access time*. Today computers are becoming available with access times on the order of 100 to 250 nanoseconds (hundred-millionths of a second). Core memories are also expensive, so that many computer systems have a limited amount of such storage and augment it with other forms that are cheaper but also slower.

A less expensive form of memory is found on a silicon chip. Although silicon memory chips are slower than core memory, they are common in small computers.

EFFICIENT OPERATION

The computer is intended to serve the computing needs of a variety of users. Even the small computer outlined in this part of the article could serve the needs of students in a computer club or a mathematics course, or the school's administrative office for certain clerical functions. Even the librarian can use it to keep track of books. Operating procedures, however, must be worked out to give each user a simple means of access to the computer.

Each user may be thought of as having a "job" to be done, and usually the user wants the job completed immediately. One procedure therefore is to allow the individ-

ual user to take over the complete computer facility to process the job. The user analyzes the problem and prepares the program to carry out the steps necessary for the computations. The user sets all of the switches and pushes all of the correct buttons on the computer console and printer to complete the computing job. Only after the job is finished is control of the facility turned over to the next user.

This mode of operation is the most inefficient and inflexible use of the computer. A long time may pass between the completion of one job and the beginning of the next, and the whole computer stands idle. Some parts of the computer may be idle for long periods of time while others are very active. For example, the printer may be idle during a long computation that keeps the CPU and storage units very busy.

If the computer is to be used almost entirely as an instructional machine, however, this "hands-on" computing experience does serve to remove some of the mystery of computing. Inefficiency is part of the cost of education. In engineering groups within even large companies, a small computer may be operated this way for the convenience of the individual engineers and their computational needs.

As more users and increasingly complex jobs begin to exert pressures on the computing facility, the need for better operating procedures becomes apparent. The work load placed upon the computer demands that the machine be active at all times, that it not stand idle while one user is finishing a job and the next user is not yet quite ready to start. The printer should not stand idle while one user is doing a very long computation. To increase efficiency as well as flexibility, the operation of the computing facility is placed under the control of an operating procedure variously known as a *monitor program*, *supervisory-control program*, *executive program*, or *operating system*. In one way or another, these programs regulate the operation of the computer. They schedule the succession of jobs on the computer, allocate the various units of the computer to different tasks, and keep

track of the accounting for machine use. They vary in size and complexity and function, but basically they remove the user from direct operation of the computer.

The simplest operating system is known as a *batch processor*. This is essentially an automated version of the manual operating procedure just outlined. The computer's work load is still a sequence of jobs, each of which commands the full attention of the computer until it is completed. Although the individual jobs are not related to each other, they form an apparently continuous stream of work by interacting with the batch-processor system, where the basic control now resides.

In batch processing, a series of jobs, often a whole day's work, is prepared in advance on some input medium. For example, with disks, the disks representing the different jobs to be done are accumulated. They are submitted to the batch processor, which brings in one job at a time, as well as the job's program and its data. When the processing and the output are completed, the batch processor brings in the next job. This continues until all the jobs have been completed.

The user no longer takes over control of the computer. The user merely submits a job to the computer center for processing, and then waits for the completed output. The expensive computing machine is kept busier doing useful work than in the hands-on environment.

The computer systems, or *hardware*, and the operating systems, or *software*, that are in everyday use are more complex than the simple computer just described. The basic principles, however, are the same.

HARDWARE AND SOFTWARE

The computer is a recent development in the long history of machines. Since the 1940's, it has emerged from the university research laboratory to become the basis of a very large industry and to serve every imaginable business and scientific application. Internal computing speeds have been reduced from one second per calculation, in the early 1940's, to less than a microsecond

in current models. Limited input and output facilities have been replaced with an almost endless variety of devices. The expensive core-storage units are augmented with magnetic disks of large capacity but slower access times.

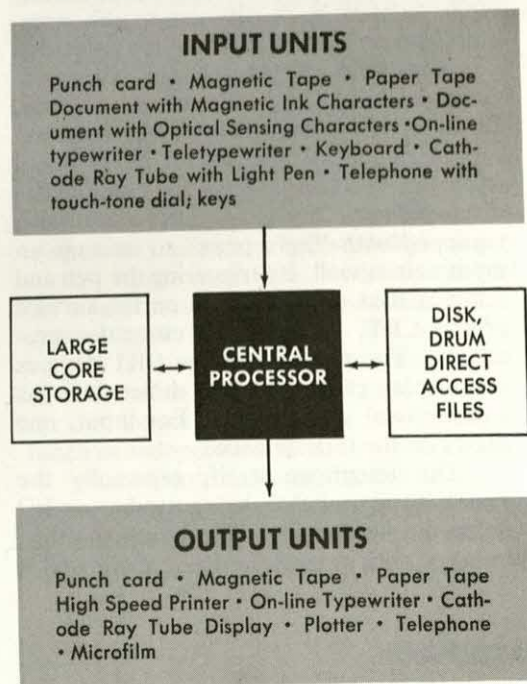
Magnetic tapes come in 800-meter reels that can pack data in densities of 625 characters per centimeter and be read or written at rates of up to 180,000 characters per second. Bank checks are printed with magnetic-ink characters that can be read automatically by the computer. Other documents are typed with a special character style that can be read by an optical-sensing device. In fact, the computer printer can print with these characters so that its output form can later be read as an input to the computer. Such forms are called *turn-around documents* and are in common use for bills from utility companies.

INPUT-OUTPUT DEVICES

The various kinds of keyboards can have their usefulness as input devices extended by making them remotely con-

This diagram shows the wide variety of input and output units that is available to computer users.

A COMPLEX COMPUTER SYSTEM





© Woodfin Camp & Assoc.

Computer-aided design (CAD) systems permit the designer to change a computer-generated image by drawing on the tube surface with a light pen.

nected to the computer. Either over a special wire connection or an ordinary telephone circuit, via a special phone coupler (modem), they are "on-line" to the central computer many kilometers away. A person typing in New York City can readily communicate with a computer in Los Angeles.

Some of the keyboard devices are specially designed for a particular application. Some are also combined with a cathode-ray tube, or CRT, display device much like a television set. Output from the computer is displayed on the CRT. CRTs are also equipped with "light pens" to become an input unit as well. By triggering the pen and pointing it at special places on the surface of the CRT, one initiates computer processing. The more extensive CRT devices will display graphs in three dimensions (as well as two) and in color. For input, one draws on the tube surface.

The telephone itself, especially the Touch-Tone model, is being used as an I/O (input-output) device. The phone buttons transmit both numeric and alphabetic codes

to the computer. The central processor is equipped to answer the phone, as a switchboard does. Computer output is translated into signals that are then transmitted back to the instrument in voice form. Thus one can dial the computer directly, send messages from the Touch-Tone panel, and receive an audible voice output.

Another output device is the plotter, which can make plots from computer-generated data. These can be point plots or line plots of great complexity and accuracy. Some plotting designs are of interest as graphic art.

COMPUTER GRAPHICS

Architects and designers are now using computers for their graphics capability. Computer-aided design (CAD) and computer-aided manufacture (CAM) are two of the tools available to computer users. Originally, CAD/CAM systems were computer simulations of two-dimensional drafting boards. More advanced systems are now able to generate three-dimensional images complete with shading and perspective. Three-dimensional simulation, or solid modeling, is used extensively in the automotive industry to reduce the need to build expensive prototypes. The motion picture industry makes use of solid modeling in film animation.

Solid modeling requires massive amounts of memory and can be done with only the largest computers. Software does most of the calculations, allowing CAD/CAM specialists and animators to concentrate on designing models.

A computer can send pictures to a CRT using a raster display, scanning the screen like a normal television picture. The screen display is divided into graphic dots. The greater the number of dots, the greater the resolution, or detail. Advanced systems have multicolored dots of variable brightness, allowing the creation of shading effects. Another system, called vector graphics, does not scan the screen continuously but uses the electron beam of the CRT to light up the objects to be drawn on the CRT screen. The rest of the screen remains blank.

DISK STORAGE

Expensive but fast core storage has been augmented by magnetic disk-storage units that have a larger capacity but slower access times for data. Disk-storage units permit large amounts of data and even computer programs to be stored at the central computer facility for access by the CPU. Some computer systems have as many as 1 billion bytes or characters of such storage.

The disk-storage unit is a rotating metal platter, or disk, that is coated with iron oxide so that it can be magnetized with spots to represent the 0 and 1 bits of a byte. These disks are mounted on a shaft that rotates at 2,500 revolutions per minute. They are packaged with several disks on a single shaft, one above another. An arm that contains two read-write heads is positioned to move in between two disk surfaces. One head serves the lower surface of the upper disk, and the other serves the upper surface of the lower disk. A whole set of such arms moves together to provide access to all the surfaces at one time, by setting at the same concentric track of each surface and moving from one track to another. These multiple arms permit access on the order of milliseconds for transfer to the core memory for processing.

Large disk files are used as permanent storage, at the computer center, for data for which almost immediate access may be needed, in contrast with tape files that are accessible only after the tapes have been mounted. Data from the disk files is also accessible without the need to search sequentially through the file as with tape. The arms may be positioned at the proper track on which the data is recorded. For this reason, such disk-storage units are often called *direct-access* storage units.

OPERATING SYSTEMS

The development of such advances in computer hardware has been paralleled by the development of suitable operating systems (software) to manage them. Let's look, as an example, at an airline-reservation system with which you have contact when you fly from one city to another.

An airline may have hundreds, and perhaps even thousands, of remote keyboard devices in its ticket offices. These keyboards are tied to a central computer facility through an extensive set of communication lines. A request for space on a certain flight is input directly to the central computer where data on all flights for several months in advance is stored on a large direct-access storage device. As a request is accepted from the ticket-sales agent, the CPU of the computer system processes the request against the flight information stored on the disk. The output, say a flight confirmation, is then transmitted back to the agent. The customer may visit a ticket office or transact business over a telephone with the agent. Either way, the agent is in direct communication with the central computer to fulfill the ticketing procedure.

One key to the development of such a reservation system is a communication network over which the thousands of terminals can transmit and receive data. A second key is the availability of large disk-storage units where data on all flights may be kept. This includes not only flight information but passenger information as well: things like the passenger's name, class of service, local telephone number, method of payment for the ticket, and so on. The third key is the operating system, the software, to make the computer system work.

The operating system must respond to an agent's request for service. It may need to service several terminals simultaneously. It must decode the incoming message and direct the computer to start the processing of programs that will read the correct flight data from the disk, enter passenger information and store the updated flight information back on the disk, and then control the transmission of the passenger confirmation back to the agent. The operating system is an elaborate control mechanism on this on-line reservation system. Its tasks are no longer so simple as controlling a batch of independent jobs.

In a different context, we can imagine a manufacturing plant that has an inventory-control system that is similar to the airline-reservation system, except that parts and

products are now the data requirement, not seats and passengers. In addition to controlling inventory, the central computer also serves a group of engineers who have small computational jobs to be done. These engineers may use a remote terminal to submit their programs and data for processing. The administrative office of the plant uses the computer for personnel applications, payroll processing, and financial forecasting.

These applications represent a variety of types of service. The inventory control represents an on-line activity that needs immediate service for updating inventory files. Some engineers would like immediate access to the machine, also on-line, while others will submit jobs to be done overnight in preparation for the next day's work. That is a batch-processing mode. The payroll office writes checks once a week or once a month. It does not need on-line action with the computer but is content with a standard type of batch processing.

The operating system must now control these different levels of activity. Multi-programming is the name given to one kind of system. The key to the operation is the partitioning of core storage so that each user has a part of the core storage for his or her own use. The operating system keeps each user program operating concurrently; that is, the computer is working on different jobs at the same time. A low-priority printing of payroll checks may go on continuously throughout the whole day, whereas engineering applications are shuffled in and out of storage as demand arises. A request for service from the inventory-control system takes priority over everything else, so that some other program may have to wait while the request is serviced. For example, the printer may stop writing checks for a few milliseconds while the spare-parts inventory file is examined.

TIME-SHARING

The current ultimate sophistication in the development of such computer-operating systems is represented by *time-sharing*. A large number of users sit at their own

terminals (keyboard or display device) on-line to the central computer. Each is entering a problem. One may be writing a program; a second debugging a program; a third sending data to a program; a fourth reading output; and so on. Each appears to have sole control of the computer as if each were the only user, much like the small hands-on computer operation. Actually, however, each user is sharing the machine with as many as two or three hundred other users.

The key to time-sharing is the fast processing capability of the machine relative to the speed with which people can use a terminal device. For example, while one user is spending five seconds inputting some data (try and see how much data you can type in five seconds), the CPU of a modern computer can execute as many as 1 million computations, or as many as 10,000 for each of 10 other users, and still not delay the person who is keying in other data. When one user is finished, the CPU processes that data while someone else is poring over some output of just thinking about what to do next.

In a time-sharing system, the user and the computer are interacting with each other in what some call a conversational mode. The user does something; the computer responds. The user does something else; and the computer responds again. This goes on in a sort of person-computer dialogue. Remote terminals in a time-sharing system may be many kilometers from the computer, anywhere that an ordinary telephone is available. With an acoustic coupler, a box that permits a typewriter or keyboard to interface with an ordinary telephone receiver, one may dial and communicate with the system. In fact, one may buy such time-sharing service from any one of a number of computer firms, paying for the use of the machine as needed and the telephone costs. As an example, without even owning a machine, a small business could use a computer by contracting for time-sharing services. In fact, several businesses could use one central computer for their own needs on such a basis.



Coleco Industries, Inc.

For many people, games have served as an introduction to electronic computers.

COMPUTER GAMES

The first commercial video game was Pong, a black-and-white simulation of a table tennis game. Its great success in 1972 created a new industry and spurred rapid development of both computer hardware and software. The first hardware developed were "dedicated" computers, capable only of playing games. The home versions of the games were played by means of a controller (joystick or paddle) and plug-in modules or cartridges containing additional memory.

Early games fell into basic categories: hit a moving object, shoot and dodge, run a maze, jump and climb. As games moved from the dedicated computer of the arcade and home video player to the personal computer, more thought-provoking games using more advanced software emerged, such as chess and adventure games.

Many games require high-resolution graphics. Newer personal computers have a special chip just to control graphic displays. Some of these chips allow variable brightness for each graphic dot in order to create shaded objects or "sprites," which are movable images such as missiles or racing cars.

Some shoot-and-dodge games require the resolution and speed of a vector graphic system. Home systems cannot use vector

graphics because it requires a specialized CRT.

Future video games will have the memory and speed to provide more realistic three-dimensional effects as well as increased user interaction. Voice synthesis and multivoiced music will also be integrated into the game format.

INFINITE NUMBER OF USES

With the development of more advanced computer hardware and software, the list of current computer applications has grown almost without limit. Computers now process voter ballots that are "marked" in some machine-readable form, such as punching out a perforated hole in a card. Computers also process bank checks on which the account numbers are preprinted in magnetic characters for automatic input. Many in the banking community are even working toward a "checkless" society. When you buy something at a store, the purchase would be recorded and the data transmitted directly to the bank computer center that serves the store. A deposit would be made to the store's account in the amount of the purchase, because the funds would be withdrawn from your account at another bank through their computer system. Not only is money not involved, not even checks are involved.



Sperry Rand Corporation



Intel Corp.

Intel Corp.



Politics is influenced by large files of data on voter characteristics, and the composition of voting districts. Basic data are updated from public-opinion polls to keep political leaders informed both about population reactions to issues as well as campaign strategies.

Data on patients in a hospital can be stored to provide faster patient histories to doctors. Monitoring patients is also possible by connecting measuring devices from the patient to a computer so that data are immediately available to the doctor to improve his medical service.

This list is indefinite, and in one sense, we are back at the beginning. The motivation behind the development of the electronic computer in the early 1940's was the need to compute the trajectories of shells for new weapons. Today the computers that grew out of the laboratory compute the orbits of earth satellites and the flight paths of probes that visit the planets. It may be fair to say that such scientific advances would not have been possible without the modern electronic computer.

Both good and evil are inherent in the vast computer potential. Large storage units serving airlines and banks also permit the collection and centralized storage of data on people. In the United States, the social security number is becoming the key to identifying its citizens. Credit data banks as well as personnel, medical, and automobile-license data banks already exist. Privacy and the theft and misuse of information are becoming issues, and debates over the uses of data are now taking place.

For this reason alone, a basic understanding of computers is necessary to an informed citizen. Only from such an understanding will you be able to help maximize the benefits and minimize the harm from one of the most fascinating of all machines—the electronic computer.

Left top: several people using a UNIVAC 9070 computer. Left middle: technicians working out on a large blueprint the intricate design for a microprocessor chip. Microprocessors are tiny silicon chips containing the complex circuits of a microcomputer's central processing unit. Left bottom: workers using microscopes assemble the chips into packages.



California Computer Products, Inc.

Mathematically programmed art is an exciting creation of the computer age. The artist feeds equations and other information into a computer to produce a design. This sand dollar design is reminiscent of a stained-glass window.

COMPUTER PROGRAMMING

by R. Clay Sprowls

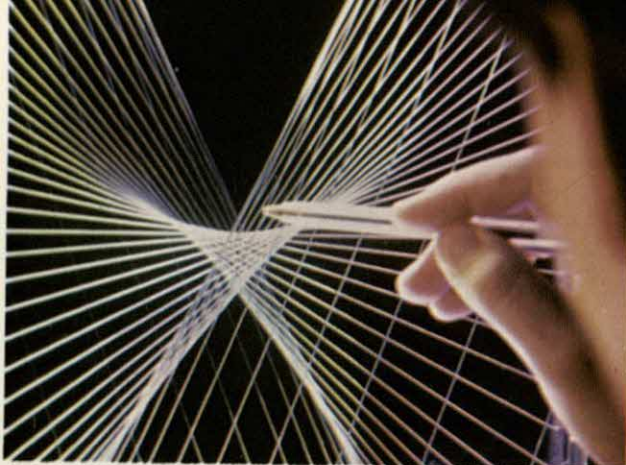
To many people, the electronic computer is a superhuman robot that can perform lightning calculations, make split-second decisions, and land a probe gently on a distant planet. True, there are computers that can do several million additions in a second, make decisions, and guide a space vehicle to a soft landing. But the computer is not superhuman; nor is it a robot, for it can do none of these things by itself.

Every computer now in existence must be told what to do: it must have a set of instructions. These instructions are called a *program*. The writing of these instructions is called *computer programming*.

A computer cannot "think" by itself. But it can perform processes that a human being can think up and program into it.

A computer is capable of obeying only a limited number of instructions. With only this limited set of instructions, human beings must describe what are sometimes very complex processes. Such descriptions often tax your mental abilities to the fullest. This is part of the fun in learning computer programming. The fun actually comes from two sources. One is devising a process that you think is correct and that the computer should be able to follow. The other is testing to see whether the process is correct and whether the computer can follow it.

The process that the computer is supposed to follow is described in a programming language. Four sample programs are shown in Table I. Each program shown performs a very simple task. Each reads



IBM

The plans for this string model of a mathematically generated surface were drawn up by a computer.

two numbers, adds them together, and prints the total. The final program, in machine language, consists of a set of binary numerals, 0 and 1, that the central processing unit of the computer can directly interpret and execute. The program is unintelligible to anyone who does not know how the computer represents data and instructions in the binary-numeral system. Contrasted with machine language, the BASIC program is almost self-explanatory.

In Table I, you can select either of two languages—BASIC or FORTRAN—that you as a programmer will use. In this table, you are adding two numbers and printing the result. BASIC and FORTRAN are examples of languages at the “user” level.

In some computers, the instructions contained in FORTRAN or BASIC must be brought down to a level closer to the computer. A language at this intermediate level is called an *assembler language*.

Finally, the instructions must be coded in binary numerals—the language the computer understands. This sequence of binary numerals is called *machine language*.

Each sample program in its own way directs the computer to follow a series of steps to accomplish the given task. The machine language is at the level of the computer itself. The others are further removed from that level and closer to a language that you can read and understand. In any of these and other programming languages, you can state precisely the process the computer is to use.

MACHINE LANGUAGE

Two devices that are important to a knowledge of programming are the central processing unit, or CPU, and the storage, or memory, unit.

The CPU is the device that executes computer instructions and controls the overall operation of the computer. The storage device is the place where data and instructions are kept for use by the CPU.

In its simplest form, a computer instruction that can be carried out by the CPU consists of two parts. One part is an operation code that gives orders. The operation code directs the CPU to establish the necessary electronic connections to carry out the order. The other part is an address of memory where the data that are to be acted upon are stored. The operation part answers the question, “What shall I do?” The address part answers the question, “Where are the data?” Machine-language instructions are coded in a form that has

TABLE I

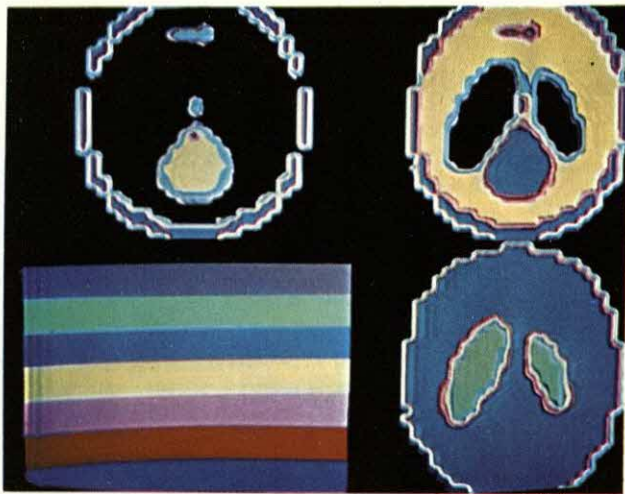
| | | |
|------------------------------|-------------------|-------------------|
| This is BASIC | 10 | READ N, M |
| | 20 | LET K = M + N |
| | 30 | PRINT K |
| | 40 | DATA 3, 5 |
| This is FORTRAN | | READ (2, 10) N, M |
| | 10 | FORMAT (I4, I4) |
| | | K = M + N |
| | | PRINT (3, 11) K |
| This is Assembly Language | 11 | FORMAT (I5) |
| | R | N |
| | R | M |
| | CLA | N |
| | ADD | M |
| | STO | K |
| This is Machine Language | PR | K |
| | 1100000000000001 | |
| | 11000000000000010 | |
| | 10000000000000001 | |
| | 10010000000000010 | |
| | 11100000000000011 | |
| | 11010000000000011 | |

been designed for a particular computer. Usually this is a binary form.

For example, the machine-language program shown has six instructions. Each instruction is a 6-digit binary numeral. The CPU of the computer takes this 16-digit binary numeral and decodes it into two parts. The first four binary digits are the operation code. The remaining 12 binary digits are the address of the data in storage. To program the computer in machine language means that you must learn the binary notation for each operation code and also how to express storage addresses by binary numerals. There may be over 200 operation codes in a large computer.

The program example can be decoded as in Table II. The accumulator that is mentioned is simply a temporary storage. The first four binary digits of the instruction uniquely identify a very simple operation that the CPU can carry out, or execute. The remaining 12 digits are the binary address of the memory location where the data are to be stored or from which they are to be fetched.

Each machine-language instruction directs the computer to take only one very small action. This may be to fetch a number from the storage unit and place it in the accumulator of the CPU, where it can be used in an arithmetic operation. It may be merely to store a number in memory, or to compare two numbers to determine if they are equal. The electronic computer gets its power and ability to solve complex prob-



Bell Laboratories

Computer versions of a brain X ray. Solid bars are colors chosen by a computer operator.

lems from the very great speed with which it can execute these elemental operations.

A complex process may need a million such small instruction steps in a program that the computer will follow. If each instruction can be executed in one-millionth of a second (a microsecond), it will take only one second to complete the entire program. Computers are able to execute instructions in much less than one microsecond, so that very complex processes represented by many millions of instructions are executed in minutes. Small processes that need only a few hundred or a few thousand instructions are executed in one millisecond (thousandth of a second) or less.

Programming in machine language is a dying art because it is too intimately involved with the machine design and with

TABLE II

| Operation Code | | Storage Location | |
|----------------|-----------------------------|------------------|---|
| 1100 | read a value | 000000000001 | into location 1 |
| 1100 | read a value | 000000000010 | into location 2 |
| 1000 | clear and add a value | 000000000001 | from location 1 into the accumulator in the CPU |
| 1001 | add a value | 000000000010 | from location 2 to the accumulator |
| 1110 | store the accumulator value | 000000000011 | into location 3 |
| 1101 | print the value | 000000000011 | stored in location 3 |

numeral systems that are not easy to handle. It is, however, the only form of program that the CPU can execute. A program written in any other form must be converted into machine language if it is to be executed on a computer.

ASSEMBLY LANGUAGE

One step removed from the CPU is programming in symbolic form, or assembly language. In assembly language, an operation code is represented by a letter or a series of letters that relates to the operation to be carried out. Thus, R may stand for READ; CLA for CLEAR AND ADD INTO THE ACCUMULATOR; STO for STORE THE CONTENTS OF THE ACCUMULATOR IN MEMORY; and PR for PRINT THE CONTENTS OF A MEMORY LOCATION. The storage locations of the data are also given letter symbols, such as N, M, and K.

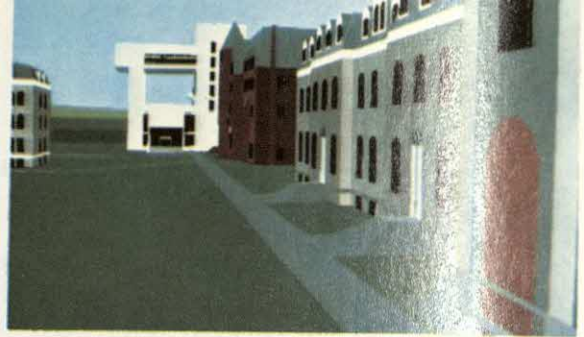
The sample assembly-language program in Table I follows the machine-language program in one-to-one order. One assembly-language instruction appears for each machine-language instruction. As a programmer, you are now concerned with symbolic operation codes and symbolic names for the storage locations, which you can devise to relate to the problem.

The CPU does not and cannot execute symbolic programming instructions. The symbolic instructions must be translated or converted into machine language. This is done automatically by the computer, using a program called an assembler. The assembler program is in machine language. Such a program is supplied with each computer when it is purchased or leased.

For many people who wish to solve problems on an electronic computer, assembly language is also too tedious and detailed. Much effort, therefore, has gone into devising higher-level languages that are further removed from the machine.

HIGH-LEVEL LANGUAGES

High-level programming languages enable you to write your instructions more nearly as they would be written in the ordinary context of the problem. Languages



A computer "walk" around Cornell University in New York. Above: Johnson Art Museum, in white, "seen" before it was built. Right: as Uris Library draws near, the perspective changes. Actually, the tower was replaced years ago.

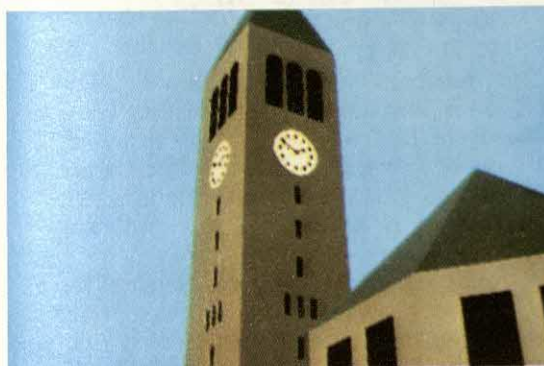
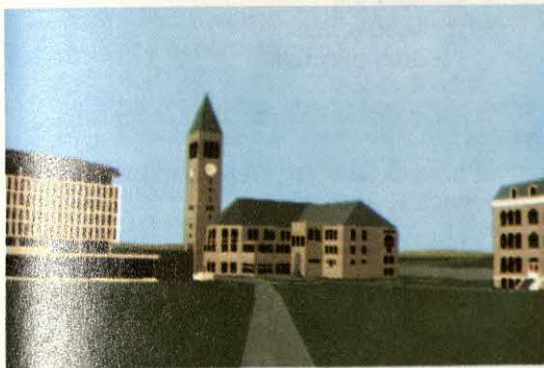
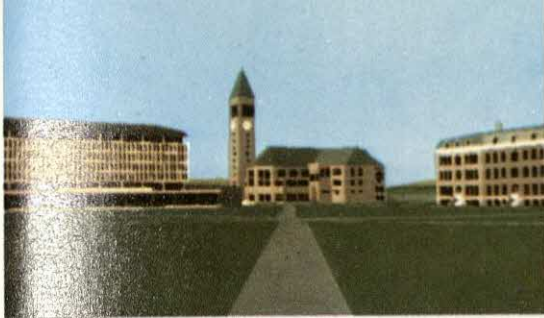
for mathematicians permit the writing of formulas. Languages for business emphasize file descriptions of business data and the kind of operations that will keep track of inventories or produce payrolls. Many languages have been devised to program a computer.

Each language is intended to make the work of programming easier. Personal computers usually have a BASIC language interpreter as standard equipment. BASIC (for Beginners All-purpose Symbolic Instruction Code) was developed in 1965 at Dartmouth College. Unfortunately, graphic control and file-handling commands vary with each manufacturer, and the BASIC language is not standardized. It is easily learned, however, and advanced versions are very useful.

LOGO is a language designed especially for children. It allows users to draw complex geometric shapes with very simple commands. It is an exploratory language with which programming and geometry are taught by trial-and-error interaction with the computer.

Other popular programming languages include Pascal, named for the mathematician Blaise Pascal, and Ada, named for Lady Ada Augusta Lovelace, also a mathematician. FORTRAN (Formula Translation) is widely used for scientific applications, and RPG (Report Program Generator) and COBOL (Common Business Oriented Language) are effective business languages.

Each language has its own rules about naming variables and writing statements to read data, do arithmetic, and print answers. Learning to program in a language is simi-



All photos, courtesy of Dr. Donald P. Greenberg, Director, Program of Computer Graphics, Cornell University in N.Y.

lar to learning any language, except that the rules are sometimes stricter than for a natural language like English.

The various programming languages, programming systems, and programs that develop around computers are called *software* to distinguish them from the physical units of the computer, or the *hardware*.

WHAT IS AN ALGORITHM?

Programming in a computer language is neither the most interesting nor the most difficult part of dealing with an electronic computer. The hardest and most interesting job is defining the problem; that is, thinking up the process that the computer is to follow.

The detailed steps that are to be followed are called an *algorithm*. It is the plan

that you propose is correct for the problem and that the computer is to follow. Once the algorithm has been written down, translating it into a programming language can actually be given to someone who knows the language but who may have no knowledge of the problem. You may find, however, that programming your own problem is part of the fun.

As an example, let's suppose that you have a homework problem in algebra that asks you to calculate one root of a quadratic equation. Such equations are a regular part of the algebra courses given in school. Solving such an equation is something you might like to use the computer for, once you have learned how to program. We shall not worry about a programming language, other than to say that you must describe in detail the individual steps that are to be executed and the order in which these steps are to be executed.

One root of a quadratic equation is expressed by

$$\text{root} = \frac{-B + \sqrt{B^2 - 4AC}}{2A}$$

When you program the solution to this computation, you are simply devising a computational procedure, or algorithm, that will direct the computer. Think about how you would explain to someone completely unfamiliar with algebra, but knowledgeable about arithmetic, how he should proceed to solve this equation. You would give him step-by-step instructions. Your instructions might go something like this:

1. Read values for A, B, and C.
2. Square the value of B and write it down.
3. Multiply A by 4 and write it down.
4. Multiply 4A (step 3) by C and record it.
5. Subtract 4AC (step 4) from B-squared (step 2).
6. Take the square root of the difference $B^2 - 4AC$.
7. Add $-B$ to the square root, and keep it.
8. Multiply A by 2, and keep it.
9. Divide the result of step 7 by 2A (step 8).

10. Write down the result of step 9 as the root.

This step-by-step procedure will lead to the computation of the root value. It is broken down into small and elemental steps of the kind that the computer can follow.

There may be one weak spot in this procedure. How do you take the square root in step 6? Do you look it up in a table? Do you get out a slide rule to compute it? Do you go to another book and learn how to compute a square root longhand? This step may need to be broken down into more detail.

In machine language or assembly language, individual instructions would be written for each of the small steps outlined, plus any additional steps to compute the square root. The whole program at this level might take as many as 50 to 75 instructions. In a high-level language, however, such as BASIC, only three statements are needed.

A BASIC PROGRAM FOR OUR PROBLEM

Three BASIC statements accomplish the square-root part of the computation: one to read the data values, one to compute the root, and one to print the answers. An additional explanation of how to take a square root is not required, because all programming languages that are defined to solve mathematics problems have basic functions built into them. To compute a square root, you write down an approved name such as SQR, and the computer will automatically carry out the steps necessary

In this frame from a TV computer animation of a tennis scene, the image has been transformed into an almost abstract color display.

Dolphin Productions



to compute the square root. This is because SQR is the name of a prepared set of machine-language instructions.

The standard programming symbols for arithmetic are:

- + for addition
- for subtraction
- * for multiplication
- / for division

** for exponentiation; that is, raising to a power

() parentheses for preferential operations, much as in algebra

A BASIC program for the root of the quadratic equation is:

```
10 READ A, B, C
20 LET ROOT = (-B + SQR (B**2
   - 4 * A * C)) / (2 * A)
30 PRINT ROOT
40 DATA 1, -4, 2
```

Line numbers indicate separate lines. READ is the statement for reading variables from the DATA line. In this example, $A = 1$, $B = -4$, $C = 2$. PRINT prints out the answer. ROOT is the variable in which the answer is stored. In this problem, $ROOT = 3.414$.

The complete program reads values for the variables, computes the root, and prints the root value. BASIC is not directly executable by the CPU, but is converted into machine-language instructions by a program called a *compiler*. The compiler program takes a single statement like that for ROOT and compiles it into many machine-language instructions. Writing in BASIC is far removed from the characteristics of the computer and more nearly like the language of the problem field—in this case, algebra.

DEBUGGING YOUR PROGRAM

Even after you have learned a programming language, you will make mistakes in writing a program. Such mistakes, or errors, are called "bugs." Locating and correcting the "bugs" is called "debugging." During the compiling of the program by the computer, the compiler has some ability to detect errors and to print out helpful hints about these errors in the form of messages called "diagnostics."

For example, you may carelessly write the statement to compute ROOT as follows:

```
20 LET ROOT = (-B + SQR (B**2  
- 4 * A * C) / (2 * A)
```

A diagnostic message will indicate a syntax error. The function SQR has its argument enclosed in parentheses as

```
SQR (B**2 - 4 * A * C)
```

and this leaves the left parenthesis in front of $-B$ without a corresponding right parenthesis, an error that the compiler can and will detect.

You can make more-subtle errors. For example, you may write the statement as

```
20 LET ROOT = (-B + SQR  
(B**2 - 4 * A * C) / 2 * A
```

The compiler will accept this as a correctly written statement, and the computer will execute it the way it is written. However, you will get the wrong answer. Now the computer does not give any hints about why the answer is wrong. Debugging becomes more difficult.

Remember that the computer will follow the directions implied in the way the equation is written. It will compute the numerator of the equation correctly. It will then divide the numerator by 2 and then multiply this quotient by A. As it is written, the programming statement accomplishes the following:

$$\text{ROOT} = \frac{-B + \sqrt{B^2 - 4AC}}{2} * A$$

The reason is simply that the product $2 * A$ is not enclosed in parentheses. Doing so indicates that the numerator is to be divided by the product. The product is the denominator of the equation. Unless you enclose $2 * A$ in parentheses, you have not properly written the direction to the computer. The computer will process what you have thought up and programmed into it. What you have thought up in this instance is well enough understood in algebra but not in computing.

You may make another kind of error. Suppose that the values of A, B, and C are respectively

$$A = 3 \quad B = 5 \quad C = 3;$$

for the BASIC program. The computer

will, in many instances, print an error message when it tries to compute the square root, because the value of the radical $B^2 - 4AC$ is -11 ; that is, it is a negative number. But there is no square root of a negative number among the real numbers.

As a programmer, you must take account of a possible negative value. You must direct the computer to compute the value as a separate quantity and then test to see if it is negative before going ahead with the calculation of the square root. The burden of doing this is on the programmer. As a matter of fact, your directions to compute the root back in the earlier section on algorithms should have included a statement about what to do when a negative value is computed, but they did not.

A FORTRAN program would be developed along similar but not identical lines. Now let us turn to a different problem to illustrate a trickier algorithm development.

ALGORITHM FOR THE FIBONACCI SERIES

An interesting series of numbers that appears in many algebra books as well as in introductory computer-science books is the Fibonacci sequence:

0, 1, 1, 2, 3, 5, 8, 13, 21, 34, . . .

In this sequence of numbers, the first two terms are given as 0 and 1. Then each term is constructed from the sum of the preceding two numbers.

You can keep on generating this series for as long as you like from the verbal explanation just given. To write a computer algorithm for this process, you must be much more explicit. A computational algorithm can be expressed by means of a flow chart, which is a graphic device for presenting the step-by-step procedure that is to be followed by the computer.

In computing the terms of this series, you quickly become aware of the fact that only two terms are used to generate the next term. Those are the "last term" and the "next to last term." After each term has been calculated, the "last term" becomes the "next to last term," and the sum becomes the "last term." This is easily seen

by moving a finger across the row of terms as the series is developed. This is also the tricky part of stating the computational algorithm.

We shall adopt the following notation, or symbols, for the terms:

N for the "next to last term"

L for the "last term"

S for the sum

A verbal flow chart is shown. The arrows indicate the direction that the steps take. The arrow that returns to the box with "Find S" forms a *loop* in the program.

After each computation of the S, the previous last term L becomes the value of the next to last term N. The sum S then becomes the value of the last term L. Then the whole process starts over again with these new values for N and L. A new S is computed. The same set of instructions is repeated over and over again in a loop.

The flow chart raises a question: How or when does the process stop? It can go on forever. The computer must be told to stop, and this is up to the programmer. You can stop after computing a certain number of terms in the series. Or you can stop

when the new term reaches a certain value. Some means or rule is needed to stop the looping computation.

Flow charts for computer programming are more formalized than the verbal one just shown. The shapes of the boxes themselves convey a meaning about the operation they contain. A small arrow indicates that a value is to be assigned to another; for example,

$N \leftarrow 0$

means to assign the value 0 to N. A diamond-shape box indicates the test of a condition. A rectangle with a curved bottom designates a printed output.

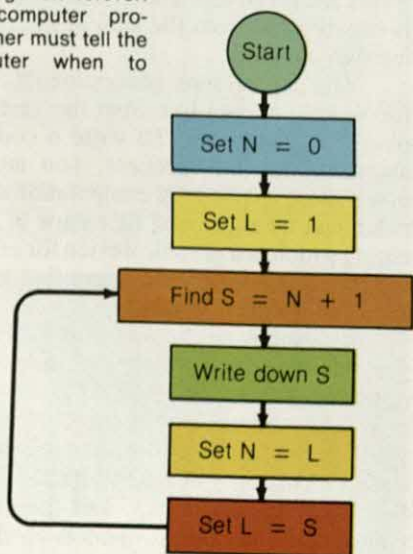
We show, on the opposite page, two flow charts for the Fibonacci series. The first one calls for a printout of each term until the term exceeds a value of 2,000. The second results in a printout only if the last term is 2,000 or less. Each is a valid computational algorithm for the series.

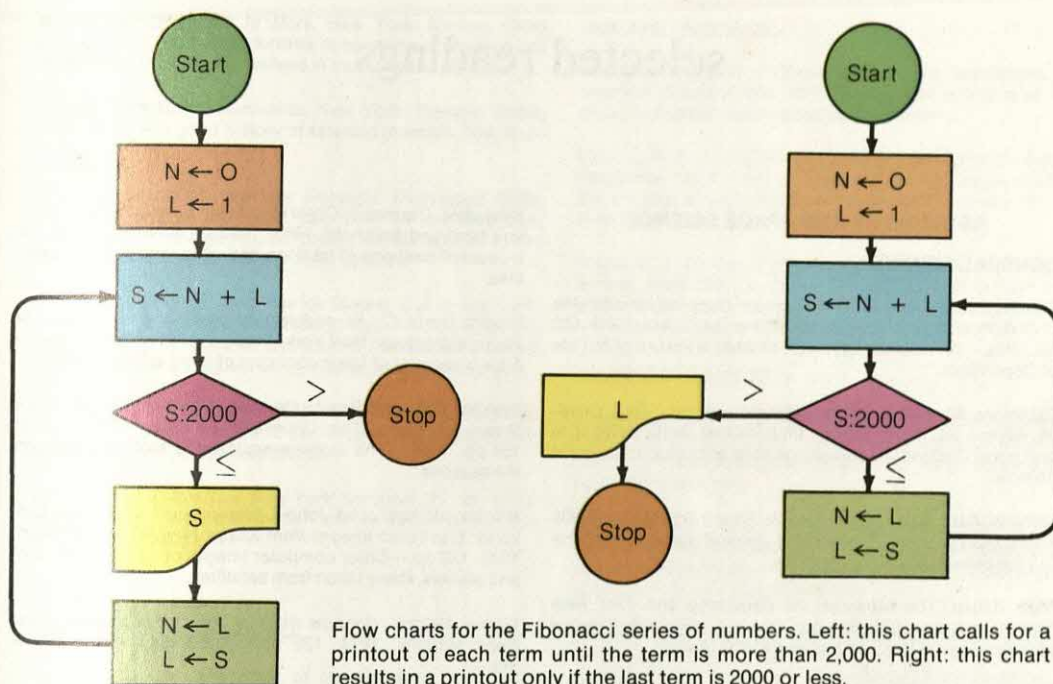
A computer program may be written directly from the flow chart in any one of several computer languages. The process the computer is to follow is completely expressed in the algorithm. The program is merely a translation of the algorithm into some language that may ultimately become executable by the computer.

MANY PROGRAMMING ACTIVITIES

Within the programmer part of computing—the software as opposed to the hardware—various levels of activity have emerged. At the lowest level is a coder who is well versed in a programming language, and who takes the algorithms that have been designed by someone else and converts them into a computer program. Higher up on the activity scale are those who work at the algorithm level. They are concerned with how to organize the solution. They may also do some coding, but this is not a requirement. Highest up are those who decide what problems to solve. For this type of work, you need some knowledge and competency in programming arts. More important is education to help you to recognize what problems need solution and to understand how the computer can play its role in the solution.

This flow chart, described in the text, could go on forever. The computer programmer must tell the computer when to stop.





More and more high schools and colleges are offering instruction in computer science that includes programming. Many books are available on programming, including instructional materials. Courses have even been developed that use the computer to teach programming. This field is sometimes called CAI for Computer-Assisted Instruction. Lessons are presented on a computer terminal, which may be as simple as a typewriter or as complex as a learning station with a typewriter, TV screen, slide projector, and tape recorder. Responses to the materials that are presented are in a dialogue with the computer; that is, with a computer program. Answers are evaluated as correct or incorrect. Directions may be given to reread a previous section or to stop and think before answering the same question again. This depends upon the instructional strategy that is programmed into the computer.

TIME-SHARING

Programming languages suitable to terminal operation and teaching are orga-

nized differently from those for which a program is submitted in its entirety. The name of this mode of computer-system operation is time-sharing. A number of users, each at his own terminal, are using a central computer simultaneously. Interactive languages are used in time-sharing systems. These may be an interactive FORTRAN or BASIC.

The essential quality of the interactive language is that as soon as a statement is entered into the computer for use in a program, that statement is evaluated for errors. Error messages appear immediately. The statement may be corrected before going on to the next statement in the program.

PROGRAMMING IS NECESSARY

Programming is necessary to computer work because the computer will carry out only a process that has been expressed in programming-language form. The computer will stand idle until programmed to do something. Only then can the computer do lightning calculations, make split-second decisions, and land a probe on Venus.

selected readings

ASTRONOMY AND SPACE SCIENCE

GENERAL WORKS

Adamczewski, Jan, et al. *Nicolaus Copernicus and His Epoch*. New York: Copernicus Society-Scribners, 1974; 160 pp., illus.—An interesting review of what is known of the life of Copernicus.

Dinsmore, Alter, et al. *Pictorial Astronomy*. New York: Crowell, 4th rev. ed., 1974; 328 pp., illus.—Clear, informative style and good illustrations provide general coverage of celestial objects.

Doherty, Paul. *Atlas of the Planets*. New York: McGraw-Hill, 1980; 143 pp., illus.—Handsome general guide for observing the planets and their satellites.

Motz, Lloyd. *The Universe: Its Beginning and End*. New York: Scribners, 1975; 345 pp., illus.—A well-written review of cosmological theories for the general reader.

Robinson, J. Hedley, and James Muir. *Astronomy Data Book*. New York: Halsted Press, 2d ed., 1979; 272 pp., illus.—Compact assortment of important astronomical data.

Sagan, Carl. *Cosmos*. New York: Random House, 1980; 365 pp., illus.—History of the universe, from the Big Bang through the evolution of human culture; a personal but fascinating view.

Van der Waerden, Bartel L., et al. *Science Awakening: II: The Birth of Astronomy*. New York: Oxford University Press, 1974; 347 pp., illus.—Why people began to study the stars and how their concepts of the universe developed and changed, through an examination of original sources. An advanced work.

Zim, Herbert S. *The Universe*. New York: Morrow, rev. ed., 1973; 64 pp., illus.—A clear and simple general discussion of astronomy; for grade school readers.

STUDYING THE SKY

Asimov, Isaac. *Eyes on the Universe*. Boston: Houghton Mifflin, 1975; 274 pp., illus.—A popular history of telescopes for the nonscientist.

Barlow, Boris V. *The Astronomical Telescope*. New York: Springer-Verlag, 1975; 213 pp., illus.—The design, engineering, and operation of modern telescopes, including space telescopes.

Donnelly, Marian. *A Short History of Observatories*. Eugene: University of Oregon Books, 1973; 164 pp., illus.—The story of the design and construction of observatories.

Howard, Neale E. *The Telescope Handbook and Star Atlas*. New York: Crowell, rev. ed., 1975; 226 pp., illus.—A good book for amateur astronomers and high school students.

King-Hele, Desmond. *Observing Earth Satellites*. New York: Van Nostrand Reinhold, 1983; 184 pp., illus.—Guide to do-it-yourself methods of tracking and observing orbiting satellites.

Knight, David C. *Eavesdropping on Space: The Quest of Radio Astronomy*. New York: Morrow, 1975; 128 pp., illus.—A clear account of basic concepts of radio astronomy.

Kyselka, Will, and Ray Lanterman. *North Star to Southern Cross*. Honolulu: The University of Hawaii Press, 1976; 160 pp., illus.—The authors explore the night sky through the seasons.

Marten, Michael, and John Chesterman. *The Radiant Universe: Electronic Images from Space*. New York: Macmillan, 1980; 128 pp.—Color computer images of galaxies, the sun, and planets, many taken from satellites.

Moore, Patrick. *Concise Atlas of the Universe*. New York: Rand McNally, 1974; 192 pp.—Good diagrams and star maps.

THE SOLAR SYSTEM

Asimov, Isaac. *The Solar System*. Chicago: Follett, 1975; 32 pp., illus.—The basic facts about the solar system; for grades 3-6.

Branley, Franklyn M. *Comets, Meteoroids, and Asteroids: Mavericks of the Solar System*. New York: Crowell, 1974; illus.—Simple and well-illustrated account, including a discussion of tectites, interplanetary dust, the solar wind, and cosmic rays.

—. *Eclipse: Darkness in Daytime*. New York: Crowell, 1973; 33 pp., illus.—A well-illustrated account for the young reader.

—. *The End of the World*. New York: Crowell, 1974; 40 pp., illus.—A simple but factual and speculative account of what may become of the earth as a planet.

Butler, S. T., and Robert Raymond Butler. *The Family of the Sun*. Garden City, N.Y.: Anchor/Doubleday, 1975; 84 pp., illus.—A panel illustration format is used to explain the nature of the solar system; suitable for grade 7 on up.

Copernicus, Nicholas. *On the Revolutions*. Baltimore: Johns Hopkins University Press, 1978; 450 pp.—Definitive English version of Copernicus's 1543 classic, with many informative notes and additional data.

Hartmann, William K., and Odell Raper. *Mars*. Washington, D.C.: NASA, 1974; 179 pp., illus.—What the Mariner voyages have revealed about Mars.

Knight, David C. *Comets*. New York: Franklin Watts, 1968; 85 pp., illus.—An excellent short, but comprehensive, study of comets.

Moore, Patrick. *Guide to Mars*. New York: Norton, 1978; 214 pp., illus.—Carefully written description of the features of Mars and the people involved in studying the planet.

Nourse, Alan E. *The Asteroids*. New York: Franklin Watts, 1975; 59 pp.—A good history of asteroid research. Well illustrated.

Short, Nicholas M. *Planetary Geology*. Englewood Cliffs, N.J.: Prentice-Hall, 1975; 361 pp., illus.—A well-illustrated description of how geological principles are applied to the study of other members of our solar system.

Tombaugh, Clyde W. and Patrick Moore. *Out of the Darkness: The Planet Pluto*. Harrisburg, Pa.: Stackpole Books, 1980; 160 pp.—Describes astronomer Tombaugh's search for Pluto, and the discoveries of Uranus and Neptune.

Washburn, Mark. *Distant Encounters*. New York: Harcourt Brace Jovanovich, 1983; 272 pp., illus.—The Voyager missions to explore Jupiter and Saturn.

Zim, Herbert S. *The Sun*. New York: Morrow, rev. ed., 1975; 64 pp., illus.—Basic information on the sun and how it is studied; for grades 6–7.

BEYOND THE SOLAR SYSTEM

Asimov, Isaac. *To the Ends of the Universe*. New York: Walker, rev. ed., 1976; 142 pp., illus.—Asimov discusses galaxies, nebulae, kinds of stars, and radio sources.

Bok, Bart J., and Priscilla E. Bok. *The Milky Way*. Cambridge, Mass.: Harvard University Press, 4th ed., 1974; illus.—The galaxy is studied from every aspect in this well-illustrated book; for the serious reader.

Cloftelder, Beryl E. *The Universe and Its Structure*. New York: McGraw-Hill, 1976; 437 pp., illus.—A good text for nonscience students.

Davies, Paul. *The Edge of Infinity*. New York: Simon & Schuster, 1983; 194 pp., illus.—Mind-bending ideas about black holes and the origin of the universe.

Kippenhahn, Rudolf. *100 Billion Suns*. New York: Basic Books, 1983; 264 pp., illus.—Nontechnical presentation of the life history of the stars.

Levitt, Ira M. *Beyond the Known Universe: From Dwarf Stars to Quasars*. New York: Viking, 1974; 131 pp., illus.—Modern astrophysics, including discussions of novae, black holes and white holes, neutron stars, and pulsars; for senior high and college.

Moore, Patrick, and Iain Nicolson. *Black Holes in Space*. New York: Norton, 1976; 128 pp.—Present theories on how stars may collapse to produce "black holes."

Sullivan, Walter. *Black Holes: The Edge of Space, the End of Time*. New York: Doubleday, 1979; 303 pp., illus.—Treats complicated topics in astrophysics in understandable terms.

Valens, Evan G. *The Attractive Universe: Gravity and the Shape of Space*. Cleveland: World, 1969; 188 pp., illus.—Well-illustrated explanation of gravitation and the way objects move in space.

TIME AND CALENDARS

Dolan, Winthrop W. *A Choice of Sundials*. Brattleboro, Vt.: Stephen Greene Press, 1975; illus.—The principle of sundials, method of communication, and history.

Elton, L. R. B., and H. Messel. *Time and Man*. Elmsford, N.Y.: Pergamon, 1979; 114 pp., illus.—Absorbing explanation of the concept of time and how it has been variously defined through history.

Krupp, E. C. *Echoes of the Ancient Skies*. New York: Harper & Row, 1983; 386 pp., illus.—A world tour of ancient temples, tombs, and observatories illustrating how the skies were used by people to create religion and calendars.

SPACE EXPLORATION

Armstrong, Neil, et. al. *First on the Moon: The Astronauts' Own Story*. Boston: Little, Brown, 1970; 434 pp., illus.—The story of the first manned landing on the moon, by the astronauts who went there.

Bova, Ben. *Workshops in Space*. New York: Dutton, 1974; 67 pp., illus.—For younger readers, a review of Skylab, earth resources technology satellites, international cooperation in space, and shuttle programs for the future.

Cortright, Edgar M., ed. *Apollo Expeditions to the Moon*. Washington, D.C.: NASA, 1975; 324 pp., illus.—Eighteen members of the NASA team write about the Apollo project.

Grey, Jerry. *Beachheads in Space*. New York: Macmillan, 1983; 274 pp.—The potentialities of space and why we must continue to explore it.

———. *Enterprise*. New York: Morrow, 1979; 288 pp., illus.—Authoritative and informative examination of the controversies behind the development of the space shuttle.

Gribbin, John. *Genesis: The Origins of Man and the Universe*. New York: Delacorte, 1981; 290 pp., illus.—Easy-to-read description of the origin of the universe and scientific answers to the fundamental questions of natural history.

Hutton, Richard. *The Cosmic Chase*. New York: NAL/Mentor, 1981; 205 pp.—Easy-to-read overview of the space race.

Kohn, Bernice. *Communications Satellites: Message Centers in Space*. New York: Four Winds, 1975; 58 pp., illus.—Concise history; for junior high readers on up.

Lewis, Richard S. *The Voyage of Apollo: The Exploration of the Moon*. New York: Quadrangle, 1974; 308 pp., illus.—Results of Apollo 12 through 17 are reviewed; for junior high on up.

Powers, Robert M. *Shuttle: The World's First Spaceship*. Harrisburg, Pa.: Stackpole Books, 1979; 256 pp., illus.—Nontechnical, readable, and well-illustrated description of space shuttles and their uses.

Stoiko, Michael. *Pioneers of Rocketry*. New York: Hawthorn, 1974; 129 pp., illus.—Review of rocketry from earliest times.

Von Braun, Wernher, and Frederick I. Ordway III. *History of Rocketry and Space Travel*. New York: Crowell, 3d rev. ed., 1975; 308 pp., illus.—The classic history in this field; for readers from junior high on up.

LIFE BEYOND EARTH

Aylesworth, Thomas G. *Who's Out There? The Search For Extraterrestrial Life*. New York: McGraw-Hill, 1975; 119 pp., illus.—Compact review of arguments for and against the possibility of interstellar communications.

Knight, David C. *Those Mysterious UFOs: The Story of Unidentified Flying Objects*. New York: Parents Magazine Press, 1975; 64 pp., illus.—For elementary students, a history of UFO reports and investigations.

Krupp, E. C., ed. *In Search of Ancient Astronauts*. New York: Doubleday, 1978; 300 pp.—The study of "archaeo-astronomy," or how ancient civilizations viewed the heavens.

Ridpath, Ian. *Messages from the Stars: Communication and Contact with Extraterrestrial Life*. New York: Harper & Row, 1978; 241 pp., illus.—The scientific background of the idea of life in other parts of the universe.

Sagan, Carl. *Communication with Extraterrestrial Intelligence*. Cambridge, Mass.: MIT Press, 1975; 428 pp., illus.—Scientists discuss the possibilities of life on other planets, and of communicating with them.

Sitchin, Zecharia. *The Stairway to Heaven*. New York: St. Martin's, 1981; 384 pp., illus.—Applies logic and scholarship to theories of ancient astronauts.

COMPUTERS AND MATHEMATICS

COMPUTERS

Freiburger, Stephen, and Paul Chew. *Consumer's Guide to Personal Computing and Microcomputers*. Rochelle Park, N.J.: Hayden, 1978; 176 pp., illus.—A guide to the lay person through the maze of available computer wares.

Harmon, Margaret. *Stretching Man's Mind: A History of Data Processing*. New York: Mason/Charter, 1975; 239 pp., illus.—From the abacus to today's computers; a semitechnical history for the advanced student.

Jefimenko, Oleg D. *How to Entertain with Your Pocket Calculator: Pastimes, Diversions, Games, and Magic Tricks*. Star City, W.Va.: Electret Scientific, 1975; 189 pp., illus.—Lighthearted approach; for grade 6 on up.

Leventhal, Lance A., and Irvin Stafford. *Why Do You Need a Personal Computer?* New York: John Wiley, 1980; 268 pp., illus.—Clearly-written, exhaustive treatment of all aspects of personal computers.

Sinclair, I. R. *Inside Your Computer*. St. Louis: Warren H. Green, 1983; 108 pp., illus.—Lucid explanations for the novice of what is inside the computer and what happens in its circuitry.

Spencer, Donald D. *Computer Dictionary for Everyone*. New York: Scribners, 1980; 191 pp.—Glossary of thousands of entries written for a wide audience.

Van Uchelen, Rod. *Word Processing: A Guide to Typography, Taste, and In-house Graphics*. New York: Van Nostrand Reinhold, 1980; 128 pp., illus.—Excellent introduction to word processing.

Willis, Jerry. *Nailing Jelly to a Tree*. Portland, Oreg.: Ilium Press, 1981; 275 pp.—Clear, logical instructions for those wanting to learn to program their own computers; emphasis on BASIC.

———, with Debra Smithy and Brian Hyndman. *Peanut Butter and Jelly Guide to Computers*. Portland, Oreg.: Ilium Press, 1978; 207 pp., illus.—Good introduction to personal computers for a wide range of readers.

MATHEMATICS

Bruno, Marilyn. *The I Hate Mathematics! Book*. Boston: Little, Brown, 1975; 127 pp., illus.—An informal presentation of a wide range of mathematical ideas; for elementary students on up.

Davis, Philip J., and Reuben Hersh. *The Mathematical Experience*. Cambridge, Mass.: Birkhauser Boston, 1981; 440 pp., illus.—Exceptional treatment of the history, nature, and significance of mathematics.

Devi, Shakuntala. *Figuring: The Joy of Numbers*. New York: Harper & Row, 1978; 157 pp.—The author's enthusiasm for numbers for their own sake communicates to the reader in descriptions of shortcuts to memory and math games.

Gowar, Norman. *An Invitation to Mathematics*. New York: Oxford University Press, 1980; 206 pp.—Makes mathematics understandable to nonmathematicians.

Hahn, James, and Lynn Hahn. *The Metric System*. New York: Franklin Watts, 1975; 63 pp.—Basic introduction; for junior high readers.

Hofstadter, Douglas R. *Gödel, Escher, Bach: An Eternal Golden Braid*. New York: Basic Books, 1979; 777 pp., illus.—Extraordinary book by a computer scientist that goes beyond any single discipline in exploring the work of the three title geniuses.

Jacobs, Harold R. *Geometry*. San Francisco: W. H. Freeman, 1974; 701 pp., illus.—A well-organized approach for senior high students.

Leonard, J. M. *Understanding Statistics*. London: English Universities Press, 1974; 216 pp., illus.—A well-written introduction to statistics for the lay reader.

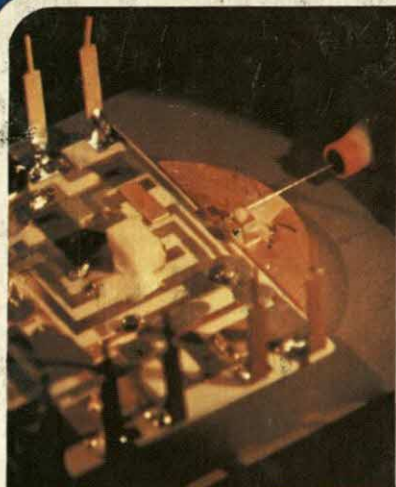
Loomis, Lynn. *Introduction to Calculus*. Reading, Mass.: Addison-Wesley, 1975; 772 pp., illus.—An intuitive approach for high school students.

Madison, Arnold, and David L. Drotar. *Pocket Calculators: How to Use and Enjoy Them*. New York: Thomas Nelson, 1978; 144 pp.—Tips on selecting calculators, their history, and how to use them.

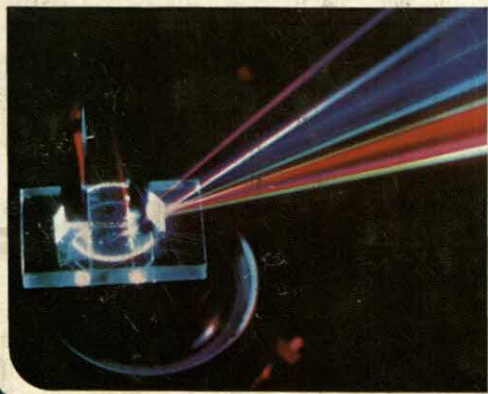
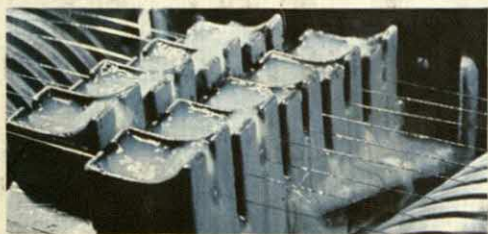
Morrison, Philip, and Phyllis Morrison. *Powers of Ten*. San Francisco: W. H. Freeman, 1983; 164 pp., illus.—An adventure in magnitudes, exploring the relative sizes of things, from quarks to galaxies.

Wells, David. *Can You Solve These?* Englewood Cliffs, N.J.: Prentice-Hall, 1983; 77 pp., illus.—Problems to test your thinking, with hints and answers.





1



ASTRONOMY
SPACE SCIENCE
COMPUTERS
MATHEMATICS
EARTH SCIENCES
ENERGY
ENVIRONMENTAL SCIENCES
PHYSICAL SCIENCES
GENERAL BIOLOGY
PLANT LIFE
ANIMAL LIFE
MAMMALS
THE HUMAN SCIENCES
TECHNOLOGY